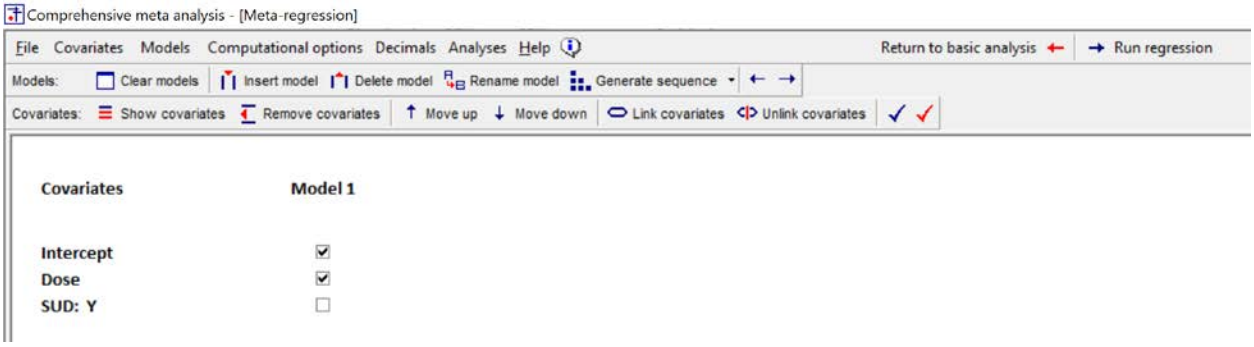
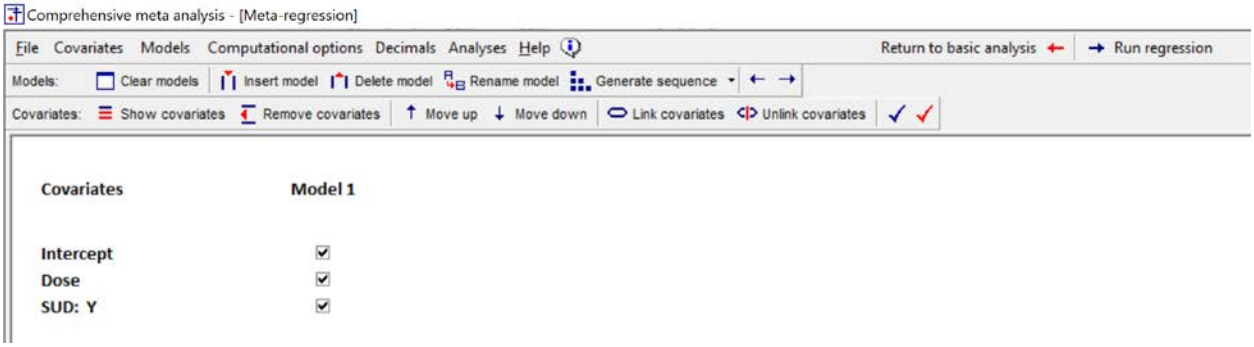
Regression in Meta-Analysis

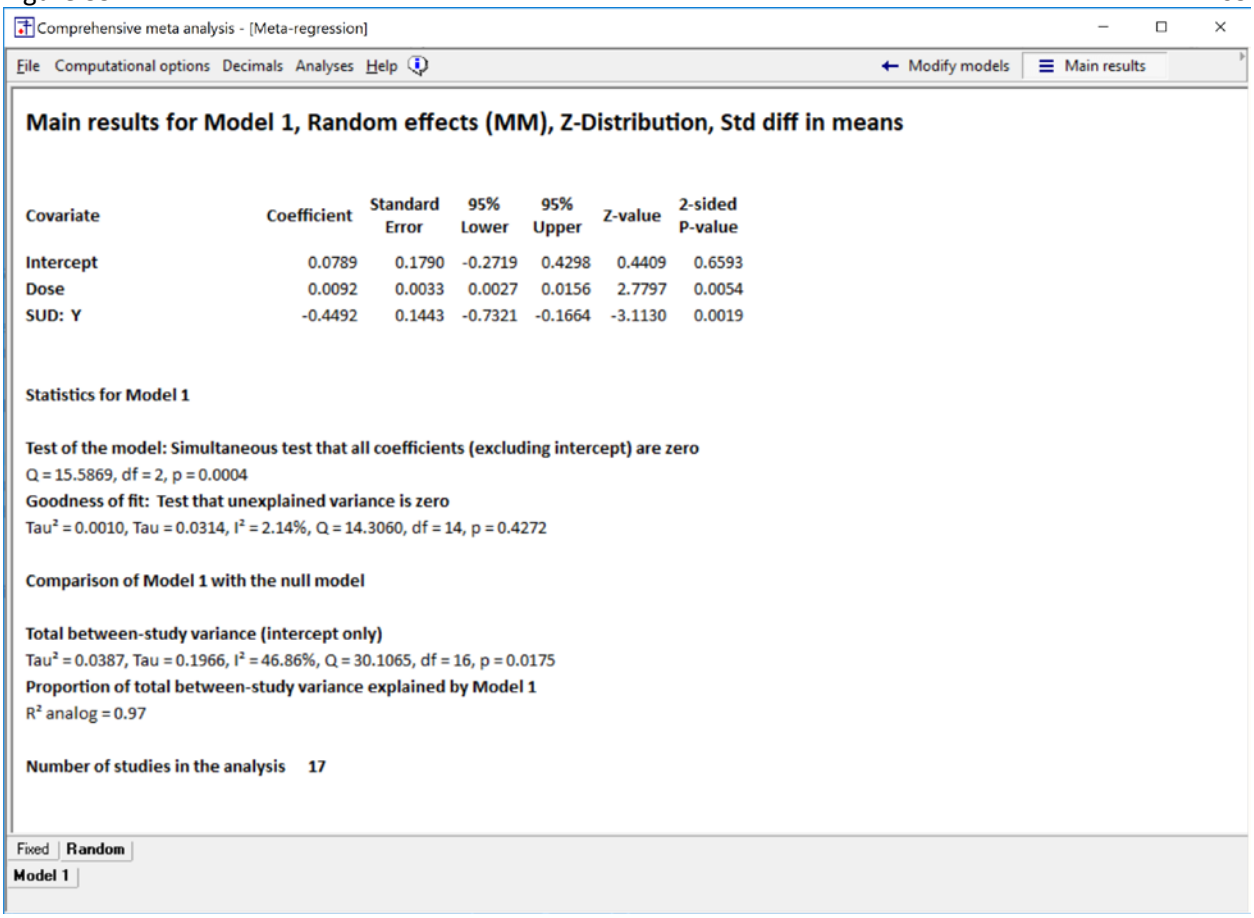
Michael Borenstein
Larry V. Hedges
Julian P.T. Higgins
Hannah Rothstein

Draft – Please do not quote

This draft edited June 8, 2017
Please send comments to Biostat100@GMail.com
An updated copy of this manual will be posted at
http://www.meta-analysis.com/pages/cma_manual.php







DATA FILES AND DOWNLOADS

This manual	http://www.meta-analysis.com/pages/cma_manual.php
CMA program	http://www.meta-analysis.com/
ADHD data in CMA format	
File using period for decimals	http://www.meta-analysis.com/downloads/ADHD P.cma
File using comma for decimals	http://www.meta-analysis.com/downloads/ADHD C.cma
ADHD data in Excel™ format	http://www.meta-analysis.com/downloads/ADHD.xls

Excel™ files for plotting interactions

http://www.meta-analysis.com/downloads/Plot-interaction-of-Hot-by-Time.xls
http://www.meta-analysis.com/downloads/Plot-interaction-of-Hot-x-Year-C.xls
http://www.meta-analysis.com/downloads/Plot-interaction-of-Latitide-x-Year-C.xls
http://www.meta-analysis.com/downloads/Plot-of-curvilinear-relationship.xls

OVERVIEW OF META-REGRESSION

INTRODUCTION

In a primary study, we may use regression analysis to study the relationship between covariates and outcome. Similarly, in a meta-analysis we may use regression to study the relationship between covariates and effect size. In this case, the procedure is sometimes called meta-regression.

In a primary study the unit of analysis is the *subject*, with covariates and outcome measured for each *subject*. In a meta-analysis, the unit of analysis is the *study*, with covariates and outcome measured for each *study*. However, with some modifications, the full arsenal of procedures that fall under the heading of "regression" in primary studies is also available in meta-analysis. For example,

- We can assess the impact of one covariate, or the combined impact of multiple covariates
- We can enter covariates into the analysis using a pre-defined sequence and assess the impact of any covariates, over and above the impact of prior covariates
- We can work with sets of covariates, such as three variables that together define a treatment, or that represent a nonlinear relationship between the predictor variable and the effect size.
- We can incorporate both categorical (for example, dummy-coded) and continuous variables as covariates.

This book is not intended as an introduction to regression analysis. We assume that the reader is familiar with the use of regression in primary studies, and our goal is to show how this procedure can be extended to a meta-analysis. We show where the procedure is the same for meta-analysis as it is for primary studies, and we also highlight the places where it's different.

While the basic ideas are the same for regression in a primary study and a meta-analysis, there are important differences in the computations. Therefore, meta-regression should only be performed using software specifically designed for this purpose. These include CMA (Comprehensive Meta-Analysis), which was developed by the authors of this book, the stata program (using the merareg macro) and R. Details on all these are included in the appendix. Most screen-shots in the text are taken from CMA.

THE ADHD EXAMPLE

We will use the "ADHD" analysis as the motivating example in this book. This analysis is reported in "Efficacy of Methylphenidate for Adults with Attention-Deficit Hyperactivity Disorder : A Meta-Regression Analysis", Xavier Castells, Josep Antoni Ramos-Quiroga, David Rigau, Rosa Bosch, Mariana Nogueira, Xavier Vidal and Miguel Casas (2001).

ADHD (attention-deficit hyperactivity disorder), a condition where people have trouble focusing on tasks, is often treated with the drug methylphenidate (brand-names include Ritalin). The meta-analysis includes studies where adult patients with ADHD were randomly assigned to receive either methylphenidate or placebo, and researchers recorded their performance on cognitive tasks.

Since different studies employed different scales, the researchers computed the standardized mean difference (*d*) between the treated and control groups, and used this for the analyses. The standardized mean difference is simply the mean difference on a standardized scale. In this example, a *d* value below zero indicates that methylphenidate was harmful, a *d* value of zero indicates no effect, and a *d* value greater than zero indicates that methylphenidate was helpful. A *d* value of 0.50 would indicate that a treated patient would (on average) score 0.5 standard deviations higher than a patient who was not treated. (For a detailed discussion of the standardized mean difference, see _____).

A simple analysis

Here, we present a simple analysis of the ADHD studies. This will serve to provide context for the regression. We will also use this example to introduce the heterogeneity statistics that we employ in meta-analysis.

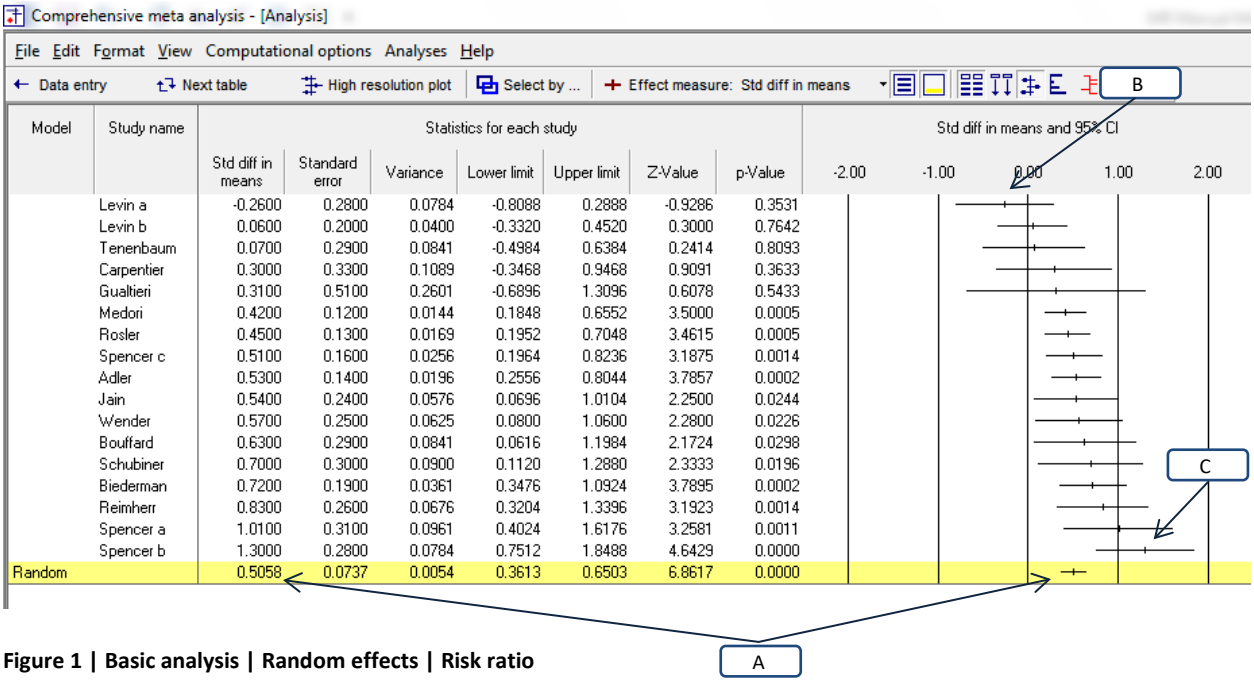


Figure 1 | Basic analysis | Random effects | Risk ratio

Figure 1 shows a random-effects meta-analysis of the ADHD studies. The *mean* effect size [A] is 0.5058 with a 95% confidence interval of 0.3613 to 0.6503, and a *p*-value of < 0.0001. Thus, there is strong evidence that methylphenidate does improve cognitive function, *on average*.

Equally important, however, is the variation in the treatment effect – how the effect size varied from study to study. The studies have been sorted in the order of effect sizes, and there seems to be substantial variation in the effects. At one extreme, there's a study [B] where the treated group did *worse* than the control group. At the other extreme, there's a study [C] where the treated group did better than the control by more than full standard deviation. And the effect sizes in the other studies fall at all points between these two extremes. Our first goal is to quantify the variation.

True effects vs. observed effects

While the issue of heterogeneity in a meta-analysis is similar to the issue of heterogeneity in a primary study, there is an important difference between the two.

In a primary study with one level of sampling we typically treat the observed scores as being identical to the true scores. We compute the standard deviation (*S*) and the variance (S^2) of the *observed* effects. These serve also as our estimates of the standard deviation and the variance of the *true* effects.

By contrast, in a meta-analysis we need to distinguish between an observed effect size and a true effect size. The observed effect size is the effect size that we see in a study. It serves as the estimate of the effect size in the study's population, but invariably differs from the true effect size in that population due to sampling error. By contrast, the true effect size is the actual effect size in the study's population. It is the effect size that we would see with an infinitely large sample size, and (it follows) no sampling error. The variance of observed effects tends to exceed the variance of true effects, and we need to take this into account when we discuss heterogeneity. Some statistics quantify variation in observed effects, some quantify variation in true effects, and some address the relationship between the two.

The heterogeneity statistics for the ADHD analysis are presented in Figure 2.

Heterogeneity				Tau-squared			
Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
30.1065	16.0000	0.0175	46.8553	0.0387	0.0310	0.0010	0.1966

Figure 2

T (at the right side of the table) is the standard deviation of *true* effects, which in this example is 0.1966. This has the same interpretation as the standard deviation in a primary study. In round numbers, the mean here is 0.50 and *T* is 0.20. If we assume that these numbers are accurate, and that the effects are normally distributed, then the true effect size in most studies will fall within two standard deviations of the mean – in the range of 0.10 to 0.90.

T^2 is the variance of true effects, which is simply the square of the standard deviation. In this example, T^2 is 0.087. This has the same interpretation as the variance in a primary study. As is true in a primary study, the variance is not terribly intuitive (since it is in squared units) but it has some statistical properties that make it useful for some analyses.

Note that T and T^2 give us the standard deviation and variance of the *true* scores, not the *observed* scores. This is critical, since it's the dispersion of the true scores that we care about.

The Q statistic is the sum of squared deviations (of observed effects from the mean) on a standardized scale. The Q statistic is computed as an interim step in computing the variance and other statistics.

The Q-value also serves for a test of the null hypothesis that the variance of true effects is zero. Equivalently, the null hypothesis asserts that true effect size is precisely the same in all studies; that the variance in observed effects is due entirely to sampling error; that if each study had an extremely large sample size (and thus trivial sampling error) the observed effects would converge on a common value.

If the null hypothesis is true, Q would be distributed as chi-squared with degrees of freedom equal to the number of studies minus 1. In our example, Q is 30.1065 with 16 degrees of freedom and a p-value of 0.0175. We reject the null and conclude that there is evidence that the true effect size does vary across studies. This test has no analog in our primary study example. There, the observed effects are treated as identical to the true effects, so (unless all scores are identical to each other) we know that the true scores vary.

Finally, we have the I^2 statistic. This statistic deals with the relationship between the variance of observed scores and the variance of true scores. The variance of observed scores incorporates the variance of true scores and the variance due to sampling error. I^2 tells us what proportion of the observed variance is due to variation in true effects rather than sampling error. Equivalently, it tells us what proportion of the observed variance would remain if each study in the analysis had an extremely large sample size and (it follows) minimal sampling error. Here, I^2 is 46.8553, which tells us that some 47% of the variance in observed effects reflects variance in true effects rather than random error variance.

There is a common belief that I^2 tells us *how much* the effects vary, but this is incorrect. I^2 is a proportion, not an absolute value. As such, it *cannot* tell us how much the effects vary, and was never intended for that purpose. We return to this issue in later chapter.

Notes.

Our goal in this volume is to explain meta-regression, and not to assess the utility of treatments for ADHD. To that end, we present analyses to illustrate aspects of regression, and do not include the full set of analyses that might be relevant to clinical decision making. Readers who are interested in the utility of the treatments should consult the original paper.

Appendix _____ shows how to download the ADHD file that we use in this example. The original analysis was based on 18 studies. However, one of those studies was missing a value for Dose, and so was excluded from the regression. In this volume, we exclude that study from all analyses for consistency.

Our intent here was to provide a conceptual introduction to the various statistics for heterogeneity. The formulas and computations for these are shown in Chapter _____.

The distribution of chi-squared under the null hypothesis as described above assumes that the error variance of the individual studies is known.

CHECK IF DOSE IS CONSTANT FOR EACH STUDY, OR A MEAN

In sum

In a primary study, we may use regression analysis to study the relationship between covariates and outcome. Similarly, in a meta-analysis we may use regression to study the relationship between covariates and effect size. In this case, the procedure is sometimes called meta-regression.

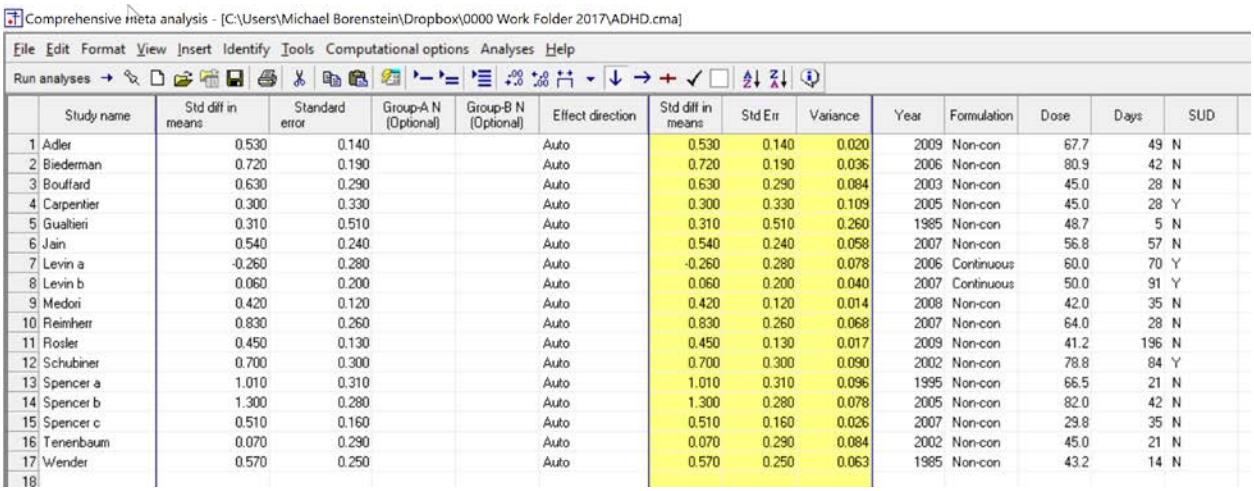
In a primary study we use estimates of the standard deviation and variance, denoted by S and S^2, to describe how the scores vary about the mean score. Similarly, in a meta-analysis, we use estimates of the standard deviation and the variance, denoted by T and T^2, to describe how the true effects vary about the mean effect.

The Q statistic is the sum of squared deviations on a standardized scale. It serves as the basis for computing the other statistics, and can be used to test the null hypothesis that the true effect size is the same in all studies. The I^2 statistic provides some context for the distribution of observed scores.

A BRIEF INTRODUCTION TO CMA

Throughout this book we will be using screen-shots from CMA (Comprehensive Meta-Analysis). Here, we provide a brief introduction to CMA. A more detailed discussion is presented in the appendix.

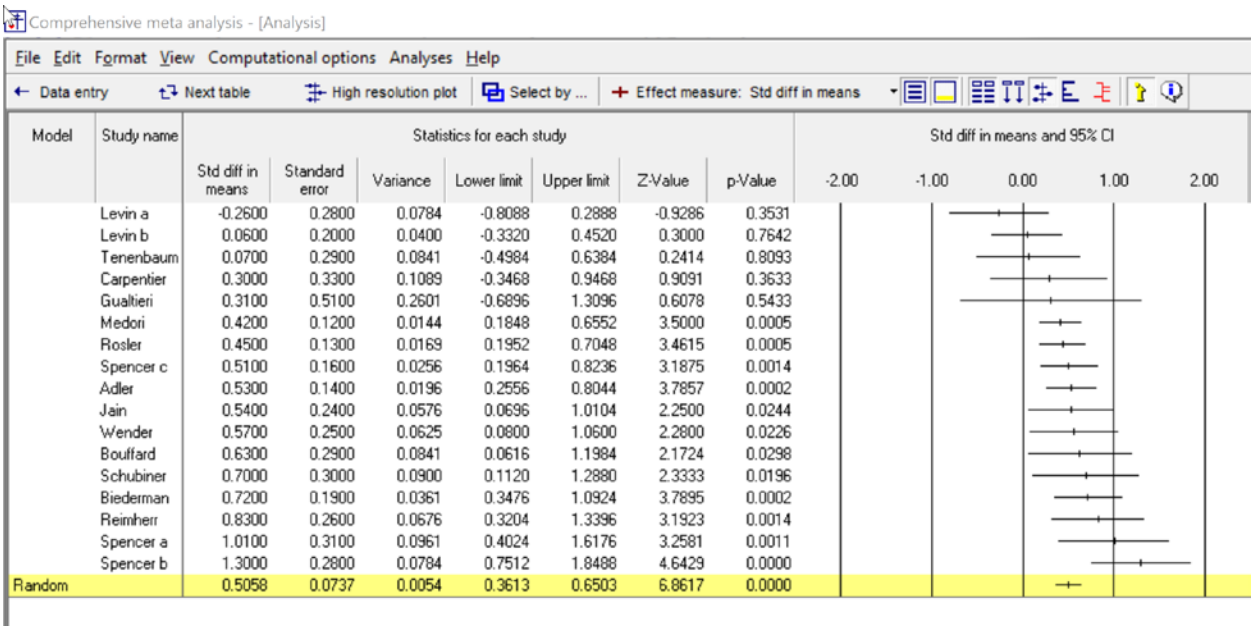
Figure 3 shows the data-entry screen in CMA. The columns at left hold the summary data for each study. The columns at right hold information about the moderators. Each moderator is defined as being either categorical, integer, or decimal. This information determines how that moderator will be treated in the regression.



	Study name	Std diff in means	Standard error	Group-A N (Optional)	Group-B N (Optional)	Effect direction	Std diff in means	Std Err	Variance	Year	Formulation	Dose	Days	SUD
1	Adler	0.530	0.140			Auto	0.530	0.140	0.020	2009	Non-con	67.7	49	N
2	Biederman	0.720	0.190			Auto	0.720	0.190	0.036	2006	Non-con	80.9	42	N
3	Bouffard	0.630	0.290			Auto	0.630	0.290	0.084	2003	Non-con	45.0	28	N
4	Carpentier	0.300	0.330			Auto	0.300	0.330	0.109	2005	Non-con	45.0	28	Y
5	Gualtieri	0.310	0.510			Auto	0.310	0.510	0.260	1985	Non-con	48.7	5	N
6	Jain	0.540	0.240			Auto	0.540	0.240	0.058	2007	Non-con	56.8	57	N
7	Levin a	-0.260	0.280			Auto	-0.260	0.280	0.078	2006	Continuous	60.0	70	Y
8	Levin b	0.060	0.200			Auto	0.060	0.200	0.040	2007	Continuous	50.0	91	Y
9	Medori	0.420	0.120			Auto	0.420	0.120	0.014	2008	Non-con	42.0	35	N
10	Reimherr	0.830	0.260			Auto	0.830	0.260	0.068	2007	Non-con	64.0	28	N
11	Rosler	0.450	0.130			Auto	0.450	0.130	0.017	2009	Non-con	41.2	196	N
12	Schubiner	0.700	0.300			Auto	0.700	0.300	0.090	2002	Non-con	78.8	84	Y
13	Spencer a	1.010	0.310			Auto	1.010	0.310	0.096	1995	Non-con	66.5	21	N
14	Spencer b	1.300	0.280			Auto	1.300	0.280	0.078	2005	Non-con	82.0	42	N
15	Spencer c	0.510	0.160			Auto	0.510	0.160	0.026	2007	Non-con	29.8	35	N
16	Tenenbaum	0.070	0.290			Auto	0.070	0.290	0.084	2002	Non-con	45.0	21	N
17	Wender	0.570	0.250			Auto	0.570	0.250	0.063	1985	Non-con	43.2	14	N
18														

Figure 3

Click "Run Analysis" to proceed to the main analysis screen (Figure 4). Each row displays the details for one study, which is also shown in the forest plot at right. The bottom row shows details for the summary effect size. A second screen (Figure 5) displays details for the heterogeneity statistics.



Model	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Levin a	-0.2600	0.2800	0.0784	-0.8088	0.2888	-0.9286	0.3531
	Levin b	0.0600	0.2000	0.0400	-0.3320	0.4520	0.3000	0.7642
	Tenenbaum	0.0700	0.2900	0.0841	-0.4984	0.6384	0.2414	0.8093
	Carpentier	0.3000	0.3300	0.1089	-0.3468	0.9468	0.9091	0.3633
	Gualtieri	0.3100	0.5100	0.2601	-0.6896	1.3096	0.6078	0.5433
	Medori	0.4200	0.1200	0.0144	0.1848	0.6552	3.5000	0.0005
	Rosler	0.4500	0.1300	0.0169	0.1952	0.7048	3.4615	0.0005
	Spencer c	0.5100	0.1600	0.0256	0.1964	0.8236	3.1875	0.0014
	Adler	0.5300	0.1400	0.0196	0.2556	0.8044	3.7857	0.0002
	Jain	0.5400	0.2400	0.0576	0.0696	1.0104	2.2500	0.0244
	Wender	0.5700	0.2500	0.0625	0.0800	1.0600	2.2800	0.0226
	Bouffard	0.6300	0.2900	0.0841	0.0616	1.1984	2.1724	0.0298
	Schubiner	0.7000	0.3000	0.0900	0.1120	1.2880	2.3333	0.0196
	Biederman	0.7200	0.1900	0.0361	0.3476	1.0924	3.7895	0.0002
	Reimherr	0.8300	0.2600	0.0676	0.3204	1.3396	3.1923	0.0014
	Spencer a	1.0100	0.3100	0.0961	0.4024	1.6176	3.2581	0.0011
	Spencer b	1.3000	0.2800	0.0784	0.7512	1.8488	4.6429	0.0000
Random		0.5058	0.0737	0.0054	0.3613	0.6503	6.8617	0.0000

Figure 4

Heterogeneity				Tau-squared			
Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
30.1065	16.0000	0.0175	46.8553	0.0387	0.0310	0.0010	0.1966

Figure 5

From the main analysis screen, click Analyses > Meta-regression 2 to proceed to the regression module. In the regression module (Figure 6), the program displays a canvas (the background) along with a list of all variables that were defined as moderators on the data-entry screen. Move any moderators onto the canvas. Tick the ones that you want to include in the model. Then click Run Regression.

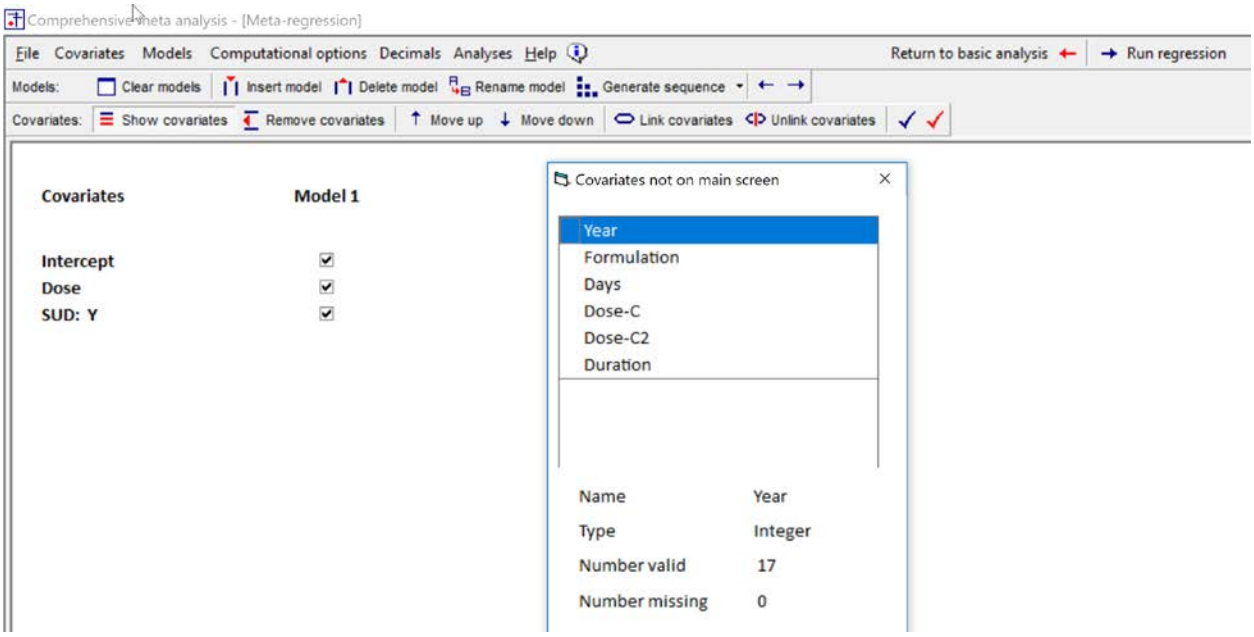


Figure 6

After you've run the regression, the program will display the results screen (Figure 7). To navigate to the plot (Figure 8) or other screens, use the menu bar.

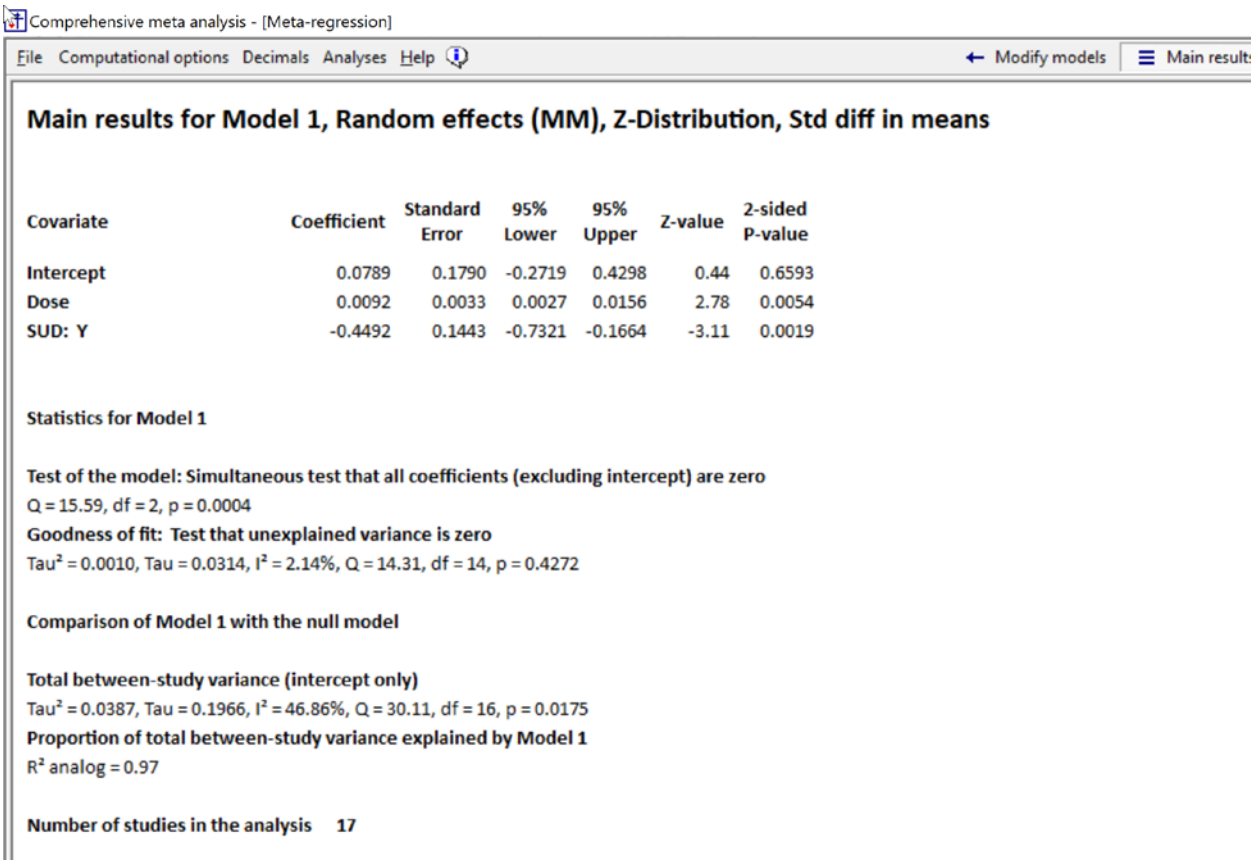

Comprehensive meta analysis - [Meta-regression]
File Computational options Decimals Analyses Help
Modify models
Main results
Main results for Model 1, Random effects (MM), Z-Distribution, Std diff in means
Covariate | Coefficient | Standard Error | 95% Lower | 95% Upper | Z-value | 2-sided P-value
Intercept | 0.0789 | 0.1790 | -0.2719 | 0.4298 | 0.44 | 0.6593
Dose | 0.0092 | 0.0033 | 0.0027 | 0.0156 | 2.78 | 0.0054
SUD: Y | -0.4492 | 0.1443 | -0.7321 | -0.1664 | -3.11 | 0.0019
Statistics for Model 1
Test of the model: Simultaneous test that all coefficients (excluding intercept) are zero
Q = 15.59, df = 2, p = 0.0004
Goodness of fit: Test that unexplained variance is zero
Tau² = 0.0010, Tau = 0.0314, I² = 2.14%, Q = 14.31, df = 14, p = 0.4272
Comparison of Model 1 with the null model
Total between-study variance (intercept only)
Tau² = 0.0387, Tau = 0.1966, I² = 46.86%, Q = 30.11, df = 16, p = 0.0175
Proportion of total between-study variance explained by Model 1
R² analog = 0.97
Number of studies in the analysis 17

Figure 7

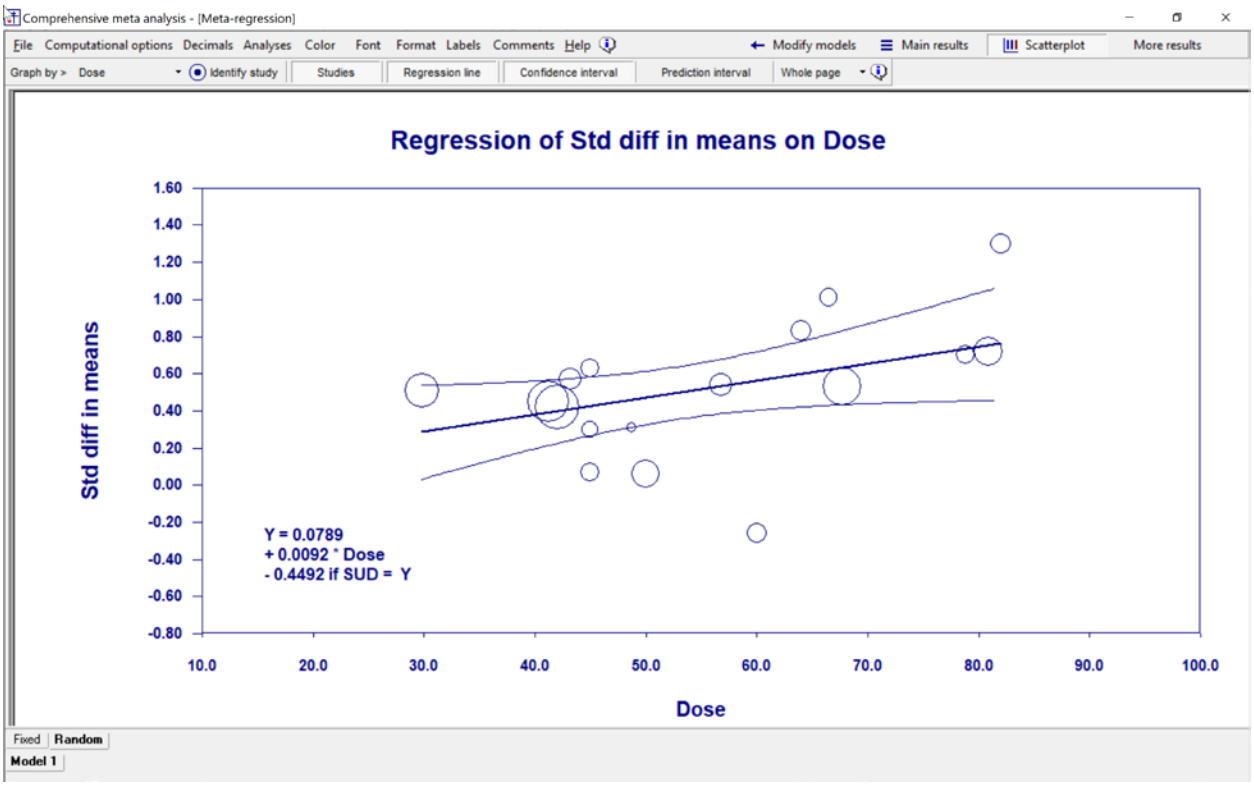


Figure 8

Other chapters in this volume provide details for the regression module, including the following.

- How to create multiple prediction models
- How to set computational options
- How to create sets of covariates
- How to work with categorical covariates
- How to include interaction terms in the prediction equation

THE ELEMENTS IN A META-REGRESSION

In this chapter, our goal is to introduce the key elements in a meta-regression. Later chapters discuss each element in detail.

In the ADHD analysis (in round numbers), the mean effect size is 0.50 and the standard deviation of true effects (*T*) is 0.20. If we assume that the true effect size in most populations will fall within some two standard deviations of the mean, then the true effect size in most populations will fall in the range of 0.10 to 0.90. This is wide range – there would be some populations where the treatment has only a trivial impact, some where it has a moderate impact, and some where the impact is very substantial. It would be important to understand why the treatment is more effective in some studies than in others. In this initial analysis, we will focus on the potential impact of the two factors, as follows.

- Dose. Was the effect size related to the dose of methylphenidate?
- SUD. Some studies exclude patients with a diagnosis of substance use disorder (SUD), while others allowed them to enroll in the study. Was this related to effect size?

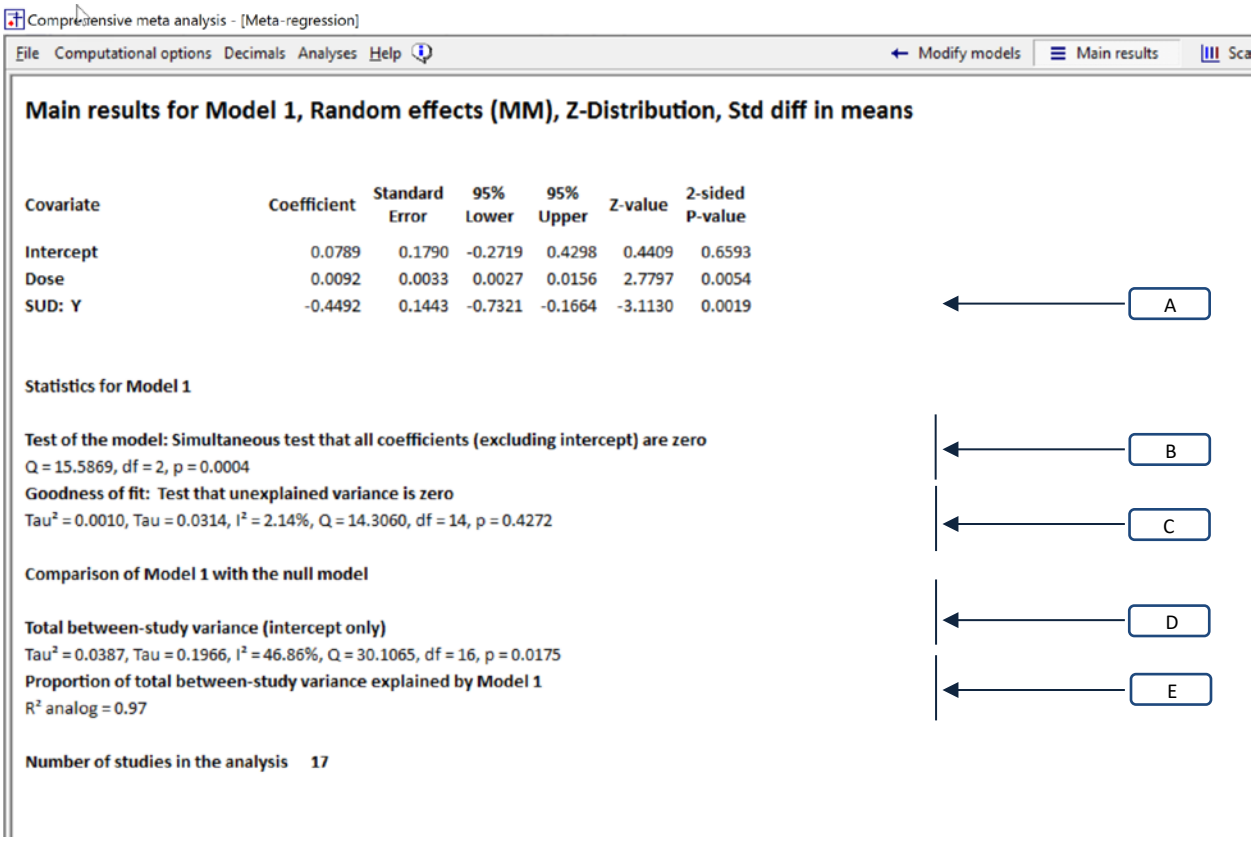


Figure 9 | Regression | Dose | Main results | Random-effects

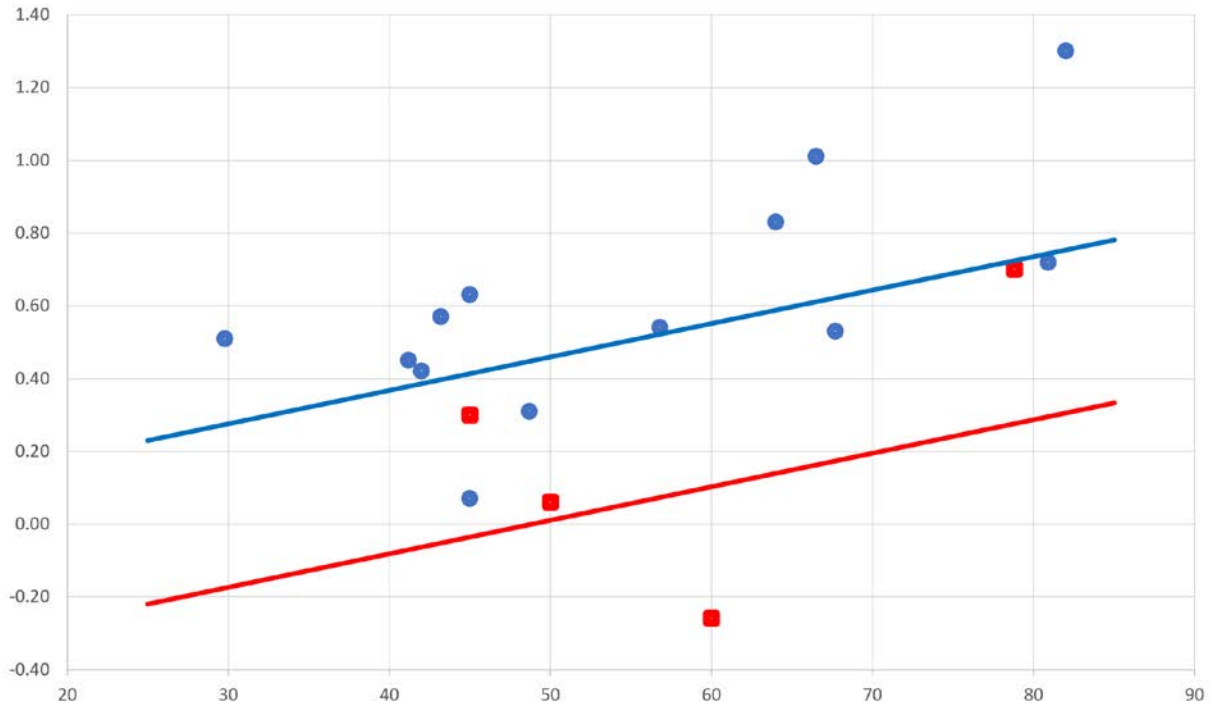


Figure 10

The results of this regression using a random-effects model are displayed in Figure 9 and plotted in Figure 10.

Test of the individual covariates (Section A)

Section [A] gives us the relationship between each covariate and the effect size, when the other covariate is held constant.

The coefficient for Dose is 0.0093, with a 95% confidence interval of 0.0027 to 0.0156. As Dose increases by one unit, the effect size increases by 0.0093 units. The *Z*-value for a test of the null (that the true coefficient is zero) is 2.7797 and the corresponding *p*-value is 0.0054. We conclude that (with SUD held constant) a higher dose probably is associated with a larger effect size. The relationship is plotted in Figure 10, where we see that as dose increases by 50 units, the effect size increases by roughly 45 points.

The coefficient for SUD (Y) is −0.4492, with a 95% confidence interval of −0.7321 to −0.1664. Studies that enrolled SUD patients had a mean effect size 0.4492 points lower than studies which excluded these patients. The *Z*-value for a test of the null hypothesis (that the true coefficient is zero) is −3.1130 and the corresponding *p*-value is 0.0019. We conclude that (with Dose held constant) the inclusion of SUD patients is probably associated with a smaller effect size. The relationship is plotted in Figure 10, where the regression line for studies that enrolled SUD patients is some 45 points lower than that for studies which excluded these patients.

Test of the model (Section B)

Where Section [A] addressed the *unique* impact of each covariate (with the other covariates held constant), section [B] addresses the *combined* impact of all covariates in the model. Here, *Q* for the model is 15.5869 with 2 degrees of freedom, and *p*=0.0004. This tells us that the model is able to explain at least some of the variance in effect size.

Goodness of fit (Section C)

In Figure 10, we see that the observed effects vary about the regression line. This variance is addressed by Section [C] in Figure 9.

The statistic T^2 is (0.0010) is the variance of true effects about the regression line. The variance is used as one element in assigning weight to the studies.

The statistic *T* (0.0314) is the standard deviation of true effects about the regression line. At any point on the regression line, if the predicted effect size is correct, then the true effect size in most populations would be expected to fall within two standard deviations of the regression line.

The I^2 statistic tells us what proportion of the variance in observed effects about the regression line is due to variance in true effects rather than sampling error. In this example I^2 is 2.14%, which means that only about 2% of the observed variance about the regression line reflects variations in true effects rather than sampling error.

The *Q* statistic (along with its degrees of freedom) is used to compute *T* and T^2. Additionally, it can be used to test the null hypothesis that all variation of observed effects about the regression lines is due to sampling error. Equivalently, the null hypothesis is that the true effect size for all studies falls directly on the regression line.

Under the null hypothesis, *Q* would be distributed as chi-squared with 14 degrees of freedom. In this example, *Q* is 14.3060 with 14 degrees of freedom and a *p*-value of 0.4272. By convention, the criterion *p*-value for this test is set at 0.10, so we cannot reject the null hypothesis. It's possible (based on this evidence) that the true effect for all studies falls directly on the regression line. This is consistent with the fact that *T* is relatively close to zero. The non-significant *p*-value suggests that the true value of *T* could be 0.0000, and that the computed value (0.0314) is simply an overestimate.

Comparison of Model 1 with the null model (Sections D and E)

Sections [D] and [E] provide information about R^2, the proportion of variance explained by the covariates. To compute R^2 we need to know the variance both with and without the covariates. Section [D] shows that the variance of true effects about the grand mean is 0.0387 (as it was in Figure 2, for the simple analysis). Section [B] showed that the variance of true effects about the regression line is 0.0010. By using these two numbers we can compute the proportion of variance explained by the covariates as 0.97.

Plot

The statistics in [A] detailed the relationship between each covariate and the effect size. The plot in Figure 10 based on these statistics.

Higher doses are associated with higher effect sizes – a 50-unit increase in Dose corresponds to a 45-unit increase in effect size. However, the absolute value of the effect size depends on whether or not the study enrolled SUD patients. For studies that excluded SUD patients, as dose increases from 30 to 80 the predicted effect size increases from (approximately) 0.30 to 0.80. For studies that included these patients, as dose increases from 30 to 80 the predicted effect size increases from (approximately) −0.20 (that is, harmful) to 0.25.

The regression line for studies that enrolled SUD patients is lower than the line for studies that excluded these patients. For any given dose, the predicted effect size for the SUD studies is around 45 points lower than that for the non-SUD studies.

Notes.

The impact of a 50-point increase in Dose (45 points) happens to be approximately the same as the impact of excluding SUD patients (45 points), but this is simply a coincidence.

In the plot, the two prediction lines are parallel to each other. Without the interaction term the lines will *always* be parallel to each other, and so this should not be taken as evidence that the impact of Dose is the same for SUD and non-SUD studies.

Summary

Each study in the analysis reported the impact of methylphenidate on cognitive scores for adults diagnosed with ADHD. The mean effect size is 0.50, but there is substantial variation in the effect size across studies. We used meta-regression to see if Dose and SUD could be used to explain some of this variance.

The model

A test of the model yields a *Q*-value of 15.5869, with 2 degrees of freedom and corresponding *p*-value of 0.0004. We conclude that the model is able to explain at least some of the variance in effect size.

Individual covariates

The coefficient for Dose is 0.0092, which means that for every one-unit increase in dose, the effect size increases by 0.0092. Equivalently, a 50-unit increase in Dose is associated with a 45 point increase in effect size. The 95% confidence interval for the coefficient extends from 0.0027 to 0.0156. The *Z*-value for a test of the null (that the true coefficient is zero) is 2.7797, and the corresponding *p*-value is 0.0054. These statistics are all for the unique impact of dose (with SUD held constant).

The coefficient for SUD|Y is −0.4492, which means that the mean effect size for studies that included SUD patients is around 0.44 lower than for studies that excluded these patients. The 95% confidence interval for the coefficient extends from −0.7321 to −0.1664. The *Z*-value for a test of the null (that the true coefficient is zero) is -3.1130, and the corresponding *p*-value is 0.0019. These statistics are all for the unique impact of SUD (with Dose held constant).

Residual variance

The variance of true effects about the regression line (T^2) is 0.00010, and the standard deviation of true effects about the regression line (*T*) is 0.0314. The I^2 statistic is 2.14%, which tells us that only about 2% of the observed variance about the regression line reflects variation in true effects rather than sampling error. The test for heterogeneity yields a *Q*-value of 14.3060 with 14 degrees of freedom and a corresponding *p*-value of 04272. We cannot reject the null hypothesis that all variance in true effects can be explained by these two covariates.

Proportion of variance explained by the model

The R^2 analog is 0.97, which means that the model is able to explain some 97% of the variance in true effects.

UNDERSTANDING THE RESULTS | RANDOM-EFFECTS

Immediately above, we introduced the statistics reported for the regression. In this chapter we will discuss each of these in detail.

TABLE OF COVARIATES (SECTION A)

In Figure 9, section [A] addresses and the relationship of each covariate with the effect size. The interpretation of these coefficients in a meta-regression is essentially the same as it would be in a primary study, as follows.

Suppose that we have two possible covariates, X_1 and X_2.

- If we include only X_1 as a covariate, the regression gives us the relationship between X_1 and the effect size, ignoring any potential confound with X_2.

- If we include only X_2 as a covariate, the regression gives us the relationship between X_2 and the effect size, ignoring any potential confound with X_1.

- If we both include X_1 and X_2 as covariates, the line for X_1 gives us the relationship between X_1 and the effect size when X_2 is held constant – the relationship between X_1 and effect size that cannot be explained as a confound with X_2. Similarly, the line for X_2 gives us the relationship between X_2 and effect size when X_1 is held constant – the relationship between X_2 and effect size that cannot be explained as a confound with X_1.

With this as context we turn to the statistics in section [A].

In our example, X_1 and X_2 correspond to Dose and SUD, respectively. The row for Dose tells us the relationship between Dose and effect size when SUD is held constant. The row for SUD tells us the relationship between SUD and effect size when Dose is held constant.

Dose

The coefficient for Dose is 0.0092. For every increase of one unit in Dose, the effect size is expected to increase by 0.0092. To make this more intuitive we can multiply both numbers by 50 – if Dose increases by 50 points, the effect size is expected to increase by some 45 points.

The coefficient reported here is an estimate of the true value in the universe of studies from which our studies were sampled. The standard error and confidence interval speak to the precision of the estimate. The standard error is 0.0033. If we assume that the true coefficient probably falls within 1.96 standard error of the estimate, then it probably falls in the range of 0.0027 to 0.0156. At the lower end, a 50-unit increase in dose would be associated with a 14-point increase in the effect size. At the upper end, a 50-unit increase in dose would be associated with a 78-point increase in effect size.

Since the range of likely true coefficients does not include zero, we can reject the null hypothesis that the true coefficient is zero, and conclude that a higher dose is associated with a higher effect size. We

can also test the null hypothesis that the true coefficient is zero. To do so we divide the coefficient by the standard error to get a Z-score of 2.7797. The corresponding p-value is 0.0054. Since the p-value is less than the criterion alpha of 0.05 we reject the null hypothesis, and conclude that a higher dose is associated with a higher effect size.

SUD

The coefficient for SUD|Y is −0.4492. The mean effect size for studies that included SUD patients is around 44 points lower than for studies that excluded these patients.

The 95% confidence interval for the coefficient extends from −0.7321 to −0.1664. The *Z*-value for a test of the null (that the true coefficient is zero) is -3.1130, and the corresponding *p*-value is 0.0019. These statistics are all for the unique impact of SUD (with Dose held constant).

The coefficient reported here is an estimate of the true value in the universe of studies from which our studies were sampled. The standard error and confidence interval speak to the precision of the estimate. The standard error is 0.1443. If we assume that the true coefficient probably falls within 1.96 standard error of the estimate, then it probably falls in the range of −0.7321 to −0.1664. The mean effect size for studies that enrolled SUD patients is at least 16 points – and possibly as much as 73 points – lower than for studies that excluded these patients.

Since the range of likely true coefficients does not include zero, we can reject the null hypothesis that the true coefficient is zero, and conclude that enrolling SUD patients is associated with a lower effect size. We can also test the null hypothesis that the true coefficient is zero. To do so we divide the coefficient by the standard error to get a Z-score of -3.113. The corresponding p-value is 0.0019. Since the *p*-value is less than the criterion alpha of 0.05 we reject the null hypothesis, and conclude that a higher dose is associated with a higher effect size.

TEST OF THE MODEL (SECTION B)

Both Section [A] and section [B] address the relationship between the covariates and effect size, but there are important differences between the two.

One difference is that section [A] tells us the direction and magnitude of the relationship between each covariate and effect size. By contrast, section [B] tells us whether or not there is evidence of a relationship, but provides no information about the nature of that relationship.

The second difference is that where Section [A] addressed the *unique* impact of each covariate (with the other covariates held constant), section [B] addresses the *combined* impact of all covariates in the model. Here, *Q* for the model is 15.5869 with 2 degrees of freedom, and *p*=0.0004. The null hypothesis here is that none of the covariates is related to effect size. The p-value is statistically significant so we reject the null, and conclude that at least one of the covariates is related to effect size.

In those analyses where there is only one covariate in the regression, sections [A] and [B] will be testing the same null hypothesis, using different (but mathematically equivalent) tests. Section [A] will report

the Z-value, while section [B] will report the Q-value, which will be equal to Z-squared. The p-value in both sections will be the same.

GOODNESS OF FIT (SECTION C)

The statistics in this section address the residual variance – the variance of effects about the regression line, or (equivalently) not explained by the covariates. To put this in perspective we will review the meaning of these statistics for a simple analysis, and then show how these meanings apply in the case of a regression.

In the case of a simple regression the predicted effect size for all studies is simply the grand mean, and so the heterogeneity statistics (Figure 2) address the dispersion of effects about this mean. The standard deviation of the true effects (T) is 0.1966. The variance of true effects (T^2) is 0.0387. The ratio of (a) the variance in true effects to (b) the variance of observed effects (I^2) is 46.86%.

These statistics are all estimates, and we can test them for statistical significance. The null hypothesis being tested is that the true values of T, T^2, and I^2 are all 0.0000. Equivalently, the null hypothesis is that the *true* effect size for all studies is precisely the same, and that all variation of *observed* effects is due entirely to sampling error.

The test statistic is Q, which (under the null) is distributed approximately as chi-squared with (k-1) degrees of freedom (where k is the number of studies in the analysis). The Q-value is 30.1065, and with 16 degrees of freedom the corresponding p-value is 0.0175. The criterion alpha for testing this null hypothesis is 0.10, so we reject the null and conclude that the true effect size varies – methylphenidate is more effective in some populations than in others.

The same interpretations apply when we run a regression.

With Dose and SUD as covariates, the predicted effect size for each study is given by the regression lines in Figure 10, and the heterogeneity statistics (Figure 9) address the dispersion of effects about this mean. The standard deviation of the true effects about the regression lines (T) is 0.0314. The variance of true effects about the regression lines (T^2) is 0.0010. The ratio of (a) the variance in true effects to (b) the variance of observed effects (I^2) is 2.14%.

These statistics are all estimates, and we can test them for statistical significance. The null hypothesis being tested is that the true values of T, T^2, and I^2 are all 0.0000. Equivalently, the null hypothesis is that the *true* effect size for all studies falls precisely on the regression lines, and that all variation of *observed* effects about the regression lines is due entirely to sampling error. As such, the null hypothesis is that the covariates are able to explain all of the variance in effect size.

The test statistic is Q, which (under the null) is distributed approximately as chi-squared with (k-p-1) degrees of freedom (where k is the number of studies in the analysis and p is the number of covariates). The Q-value is 13.3060 with 14 degrees of freedom, and the corresponding p-value is 0.4272. The criterion alpha for testing this null hypothesis is 0.10.

If the observed *p*-value was less than 0.10 we would reject the null, and conclude that the covariates do no explain all the variance in effect size – even if we know the Dose and SUD status for a study, we will not know the true effect size for that study precisely. In our example the *p*-value is 0.4272 and so we cannot reject the null. Based on this evidence, it's possible that if we know the Dose and SUD for a study, we can predict that true effect size for that study with no error.

UNDERSTANDING T^2, T, Q

When we run a regression in a primary study, the statistics that quantify the heterogeneity are the standard deviation (S), the variance (S2) and the sum of squared deviations (SS). The relationship among these statistics is reasonably straightforward. Our goal in the regression is to explain some of the variance (S2). In the regression we work with the sum of squares (SS), which allows us to perform some statistical tests. Finally, if we want to know how widely the scores vary we use the standard deviation (S). The standard deviation is simply the square root of the variance, and allows us to work on a linear scale.

The same basic idea applies also to regression in a meta-analysis. The statistics that quantify the heterogeneity are the standard deviation of true effects (T), the variance of true effects (T2), and the sum of squared deviations on a standardized scale, Q. Our goal in the regression is to explain some of the variance (T2). We work with the sum of squares (Q), which allows us to perform some statistical tests. Finally, if we want to know how widely the effects vary we use the standard deviation (T). The standard deviation is simply the square root of the variance, and allows us to work on a linear scale.

The two paragraphs above highlight two key points.

The goals and the statistics for regression in a meta-analysis are analogous to those in a primary study. Concretely, T2 in the meta-analysis is analogous to S2 in the primary study. Q in the meta-analysis is analogous to SS in the primary study. And T in a meta-analysis is analogous to S in a primary study.

At the same time, there are some important differences in the way these statistics are computed. These stem from the fact that in a primary study with one level of sampling we typically treat the observed scores as being identical to the true scores, whereas in a meta-analysis we distinguish between the two. Therefore, in a meta-analysis, we have some statistics that address the heterogeneity in observed scores, some that address the heterogeneity in true scores, and some that address the relationship between the two. In the appendix we show how these statistics are related to each other, and how each is computed. Here, our goal is to outline what each statistic means, and (thus) which ones we should focus on when interpreting the results.

There is one statistic reported in a meta-analysis that does not have an analog in a primary study. That statistic is I2. Because I2 is widely misused, we discuss that statistic in its own section, immediately below.

UNDERSTANDING I^2

As of this writing, the statistic most often cited to quantify variation in effects is the I2 statistic. It is widely believed that I^2 tells us how much the effect size varies across studies. In fact, this interpretation is incorrect – it represents a fundamental misunderstanding of this statistic. I^2 is a proportion, not an absolute value. It tells us what proportion of the variance in observed effects reflects variance in true effects rather than sampling error. It does not tell us how much the true effects vary.

We can reinforce this point with a simple thought experiment. Suppose we are told that a meta-analysis has a mean effect size of 0.50, and that I^2 is 50%. Then we're asked how much the true effect size varies across studies. Does it vary from 0.40 to 0.60, or from 0.30 to 0.70, or across some other range? The fact of the matter is, there is no way to know. I^2 was never intended to capture this kind of information.

With that out of the way, we can discuss the meaning and the correct use of this statistic.

Recall that the variance of observed effects incorporates the variance of true effects, and also the variance due to sampling error. I^2 tells us what proportion of the variance of observed effects is due to variance in true effects rather than sampling error. As such, it also tells us what proportion of the observed variance would remain if we could somehow remove the sampling error and plot the true effects rather than the observed effects.

A conceptual definition of I^2 would be the ratio of true variance to total variance, which could be expressed as

$$I^2 = \frac{V_{TRUE}}{V_{OBS}} = \frac{V_{TRUE}}{V_{TRUE} + V_{ERR}} = \frac{T^2}{V_{OBS}} \tag{1.1}$$

The computational formula is

$$I^2 = \frac{Q - df}{Q} \tag{1.2}$$

In this formula, $Q - df$ is proportional to V_{True}, while Q is proportional to V_{OBS}.

For example, in the ADHD analysis the variance of observed effects is _____. We don't generally report this number, but we've done so for the purpose of this example. And, as reported earlier, the variance of true effects is _____. The ratio of one to the other gives us an *I2* value of 47%.

If that's the meaning of I^2, what role does it play in an analysis? It provides some context for the forest plot, and this context is captured by the formula

$$T^2 = V_{OBS} \times I^2 . \tag{1.3}$$

This formula is the key to using I^2 correctly. I^2 provides context for the observed effects – if we multiply the variance of observed effects by I^2, we get the variance of true effects. In practice, if we are looking at a plot of the observed effects, then I^2 tells us what proportion of the observed variance would remain if we could somehow remove the sampling error from the plot. If I^2 is high, this tells us that most of the observed dispersion would remain. If I^2 is low, this tells us that most of the observed dispersion would disappear.

To this point we've explained what it means in Section [D] for the simple analysis, with no covariates. Here, I^2 is 47% and it tells us that about half the variance in _____ reflects variation in true effects, and would remain if we could somehow remove the sampling error.

Precisely the same idea applies in Section [C] for the regression with Dose and SUD as covariates. Here, I2 is around 2% and it tells us that only about 2% of the variance in _____reflects variation in true effects, and would remain if we could somehow remove the sampling error.

If we do want to use I^2, this would be the way to use it. However, a more accurate approach is to simply estimate the range of true effects using T, as discussed in the section on prediction intervals.

As we've noted above, it's common for researchers to interpret I2 (incorrectly) as telling us how much the effects vary. By the same logic, researchers will sometimes use the change in I2 (from section D to C) to tell us how much variance is explained by the covariates. Again, this is a complete mistake, since I2 does not tell us how much variance there is.

If we look at the change in I2, we are saying that "Initially, 47% of the observed variance was due to variance in true effects, and now 2% of the variance is due to variance in true effects". There's no reason to care about this, and this is not what the researcher thinks is being measured.

Summary

If we start with a plot of the effects (either about the grand mean or about the regression line), some of the observed dispersion reflects differences in the true effects, while some reflects sampling error.

We start with the dispersion of effects about the grand mean. I^2 tells us what proportion of this reflects variation in true effects, and as such may be explained by study-level covariates.

When we add covariates we are looking at the dispersion of effects about the regression line. I^2 tells us what proportion of this reflects variation in true effects, and as such may be explained by additional study-level covariates.

In both cases, what we really care about is the absolute amount of variation (not the proportion) and this is captured by T^2 or T. These values tell us if the amount of dispersion is of substantive or clinical importance.

The role of I^2 is to provide context for the dispersion that we see in the plot.

- If I^2 is near 100%, we know that if the studies had extremely large samples sizes (trivial error) the pattern of scatter would remain largely unchanged.
- If I^2 is near 0%, we know that if the studies had extremely large samples sizes (trivial error) the effects would converge on the regression line.

CONFIDENCE INTERVALS AND PRECISION INTERVALS

Our goal in this chapter is to explain what confidence intervals are, and what prediction intervals are. Researchers sometimes confuse the two with each other, but in fact they are two entirely different things. We will start with a simple case – a primary study, where our goal is to estimate the mean score – and use that to establish the meaning of these statistics. Then we will show how the same idea applies in a simple meta-analysis and in a meta-regression.

A primary study

Consider a primary study where our goal is to determine how students' scores are distributed in a particular school. We draw a random sample of students, and record their scores.

Often, we look first to the mean, which tells us how well the students are performing *on average*. A class with a mean score of 70 is obviously different than one with a mean score of 30 – the average student in the former is doing better than the average student in the latter. At the same time, we need to consider also how the student scores vary about the mean. A class with a mean of 50 with scores that vary from 10 to 90, is very different than a class with a mean of 50 with scores that all fall in the range of 40 to 50.

The confidence interval tells us how *precisely* we've estimated the mean. The confidence interval may be estimated as covering roughly two *standard error* on either side of the mean. If the mean in our *sample* is 50 and the standard error is 5, then the confidence interval is roughly 40 to 60. This tells us that the mean in the *population* probably falls somewhere in this range.

The prediction interval tells us how narrowly (or widely) the scores are *dispersed* about the mean. The prediction interval may be estimated as roughly two *standard deviations* on either side of the mean. If the population mean is 50 and the standard deviation is 20, then the prediction interval is roughly 10 to 90. This is called a prediction interval because if we were asked to predict the score for any one student (selected at random from the class) we would predict that student would score in the range of 10 to 90. And we'd be correct some 95% of the time. From a substantive perspective this is a fairly wide interval. Some students are performing poorly, some are performing at grade level, and some are performing exceptionally well.

Note that the confidence interval is a property of the *sample* – It tells us how precisely we've estimated the mean score for all students in the population. If we increase the sample size, the confidence interval will narrow. With a large enough sample size, the width of the confidence interval would approach zero and we would know the population mean with only a trivial amount of error.

By contrast, the prediction interval is a property of the *population*, and reflects how the score are actually distributed. If the scores vary from 10 to 90, then they vary from 10 to 90 whether we elect to draw a sample of 10 students, or 20, or 50. As researchers we may have the power to get a better estimate of a parameter, but not to change how the students are actually performing.

In sum, we might report that the mean score is 50 with a confidence interval of 40 to 60 and a prediction interval of 10 to 90. The confidence interval is an index of precision, and tells us that the

mean probably falls in the range of 40 to 60. The prediction interval is an index of dispersion, and tells us that some students score as low as 10 and others score as high as 90.

A simple meta-analysis

The same ideas apply in a meta-analysis.

Our sample of 17 studies has been sampled from a universe of populations, and we are trying to estimate the mean effect size in that universe.

The confidence interval tells us how precisely we've estimated the mean. The confidence interval may be computed as roughly two *standard error* on either side of the mean. In the ADHD example (in round numbers) the mean is 0.50 and the standard error is 0.074, so the confidence interval is roughly 0.35 to 0.65. The true mean (for the universe of all relevant populations) probably falls somewhere in this range.

The prediction interval tells us how narrowly (or widely) the effects in different populations are dispersed about the mean. The prediction interval may be computed as roughly two *standard deviations* on either side of the mean. The mean is 0.50 and the standard deviation is 0.20, so the prediction interval is roughly 0.010 to 0.90. In most populations (drawn from the same universe as the populations in the analysis) the true effect size will fall in this range. From a substantive perspective this is a wide range – in some populations the impact of the treatment will be trivial, in some it will be modest, and in some it will be striking.

Note that the confidence interval is a property of the *sample* – It tells us how precisely we've estimated the mean score. If we increase the sample size, the confidence interval will narrow. With a large enough sample size, the width of the confidence interval would approach zero and we would know the mean with only a trivial amount of error.

By contrast, the prediction interval reflects how the treatment effects score are actually distributed. If the effect sizes vary from 0.10 in some populations to 0.90 in others, then (across all populations in the universe) they vary from 0.10 to 0.90 – whether our analysis includes a sample of 10 studies, or 20, or 50. As researchers we may have the power to get a better estimate of a parameter, but not to change how the treatment actually works in different populations.

In the ADHD example (Figure 11), the mean score is 0.50 with a confidence interval of 0.35 to 0.65 and a prediction interval of 0.10 to 0.90. The confidence interval [A] is an index of precision, and tells us that the mean effect probably falls in the range of 0.35 to 0.65. The prediction interval [B] is an index of dispersion, and tells us that the effect size will be as low as 0.10 in some populations and as high as 0.90 in others.

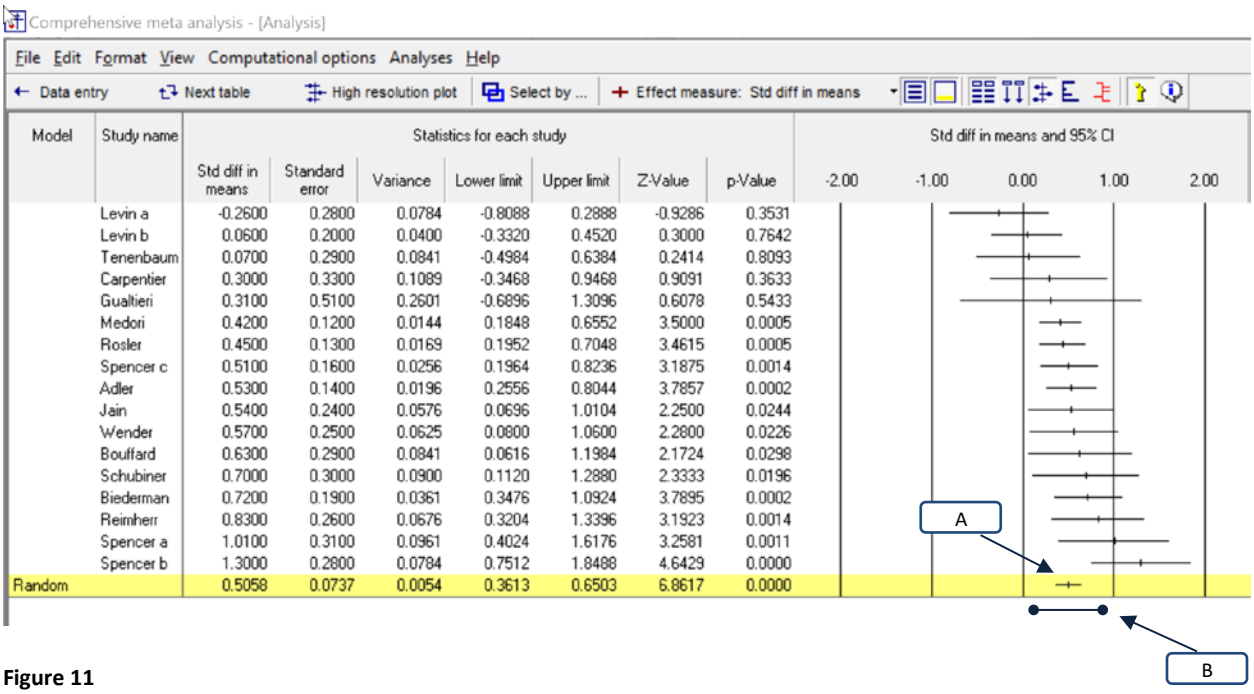
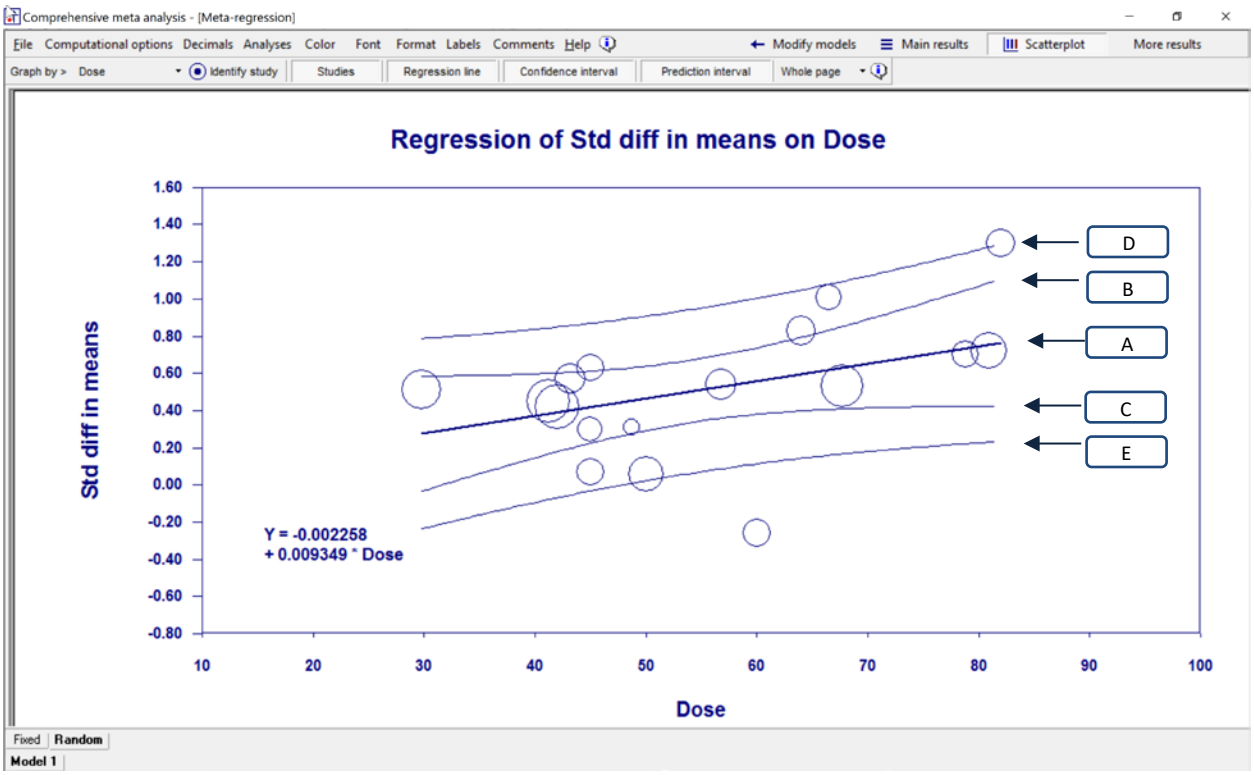


Figure 11

Meta-regression

Finally, the same idea applies to regression. Consider the plot shown here, which shows the predicted effect size as a function of Dose. The regression line is bounded by two pairs of lines – one shows the confidence interval, and the other shows the prediction interval.



For a study where the Dose is 80, the predicted effect size (in round numbers) is 0.75 [A].

The confidence interval is approximately 0.40 to 1.10 [B to C]. If we take all studies in the universe with a Dose of 80, the true *mean* for all these studies probably falls in the range of 0.40 to 1.10. But that tells us only about the mean – it says nothing about the dispersion of effects about the mean. The prediction interval is 0.20 to 1.20 [D to E]. If we look at all these studies (with a Dose of 80) the true effect size for most of them will fall in the range of 0.20 to 1.30. Or (equivalently) if we choose one of these studies at random, it will probably have a true effect size in this interval.

Note that this is the same interpretation as before – the only difference is that before, we had one interval for all studies, and now we have an interval based on the Dose.

Context

In a simple analysis we typically ask if the mean effect is clinically important, but we need to look at the mean in the context of the entire distribution of effects. If all effects fall within a narrow interval of the mean, then the mean may serve as a useful index for reporting the impact's utility. By contrast, if the effect size varies substantially across populations, then the potential utility of the intervention depends not only on the mean, but also on the range of effects. By looking at the full range of effects we can determine (for example) if the treatment is (a) harmful in some cases and helpful in others, (b) trivial in some cases and moderately helpful in others, or (c) moderately helpful in some cases and very helpful in others. This is the information captured by the prediction interval.

Computations

In keeping with the goals of this chapter, we suggested that the confidence interval can be computed using the mean plus or minus two standard error, while the prediction interval can be computed using the mean plus or minus two standard deviations. This simple formulation highlights the difference between the two, but ignores a number of issues that we need to address in the computations.

The simple formulas outlined above will work if the effect size index is a raw difference in means, a standardized difference in means, a risk difference, or any other index where the computations are carried out in the original metric. By contrast, the formula would not work for a risk ratio. When we are working with a risk ratio the mean effect is reported as a risk ratio but the standard error and standard deviation are reported in log units. In this case we would transform the mean effect into log units, compute the intervals, and then convert them back into risk ratio units for reporting. The same applies to other indices where the analysis takes place in log units or some other transformed metric. The computations are typically handled by the software.

The formulas assume that the mean, standard error, and standard deviation are known. And, the formulas will work reasonably well when these values are estimated with good precision. However, when the estimates are imprecise, the formulas will not work reliably. The formulas can (and should) be modified to take into account the fact that we are working with estimates. When the analysis includes a reasonable number of studies, the adjusted intervals may be a little wider than the naïve values. However, when the analysis includes only a small number of studies, the adjusted intervals could be substantially wider than the naïve values. The adjusted values should be used as matter of course. Formulas that include these adjustments are discussed in the appendix. They are also built into the program, and can be invoked by selecting the Knapp-Hartung adjustment.

UNDERSTANDING THE R^2 ANALOG

When we run a regression in a primary study we typically report a statistic called R^2, which is the proportion of variance that is explained by the covariates. In theory, R^2 can range from 0.00 to 1.00, though in practice the actual range encountered is somewhat more limited. The R^2 statistic is useful since it serves as a kind of effect size for the model – it provides some sense of how well the model is able to explain the variance, and allows us to evaluate the current model versus other models.

If V_{Total} is the variance of all scores about the grand mean, $V_{Residual}$ is the variance of all scores about the regression line, and $V_{Explained}$ is the variance explained by the model, then we can compute R^2 using

$$R^2 = \frac{V_{Explained}}{V_{Total}}, \tag{1.4}$$

Or equivalentyly,

$$R^2 = 1 - \left(\frac{V_{Err}}{V_{Total}} \right). \tag{1.5}$$

Similarly, when we use regression in a meta-analysis we can report the proportion of variance in true effects explained by the covariates, called the R^2 analog, and computed as

$$R^2 = \frac{T^2_{Explained}}{T^2_{Total}}, \tag{1.6}$$

or equivalently as

$$R^2 = 1 - \left(\frac{T^2_{Err}}{T^2_{Total}} \right). \tag{1.7}$$

This is called the R^2 *analog* rather than R^2 because there is a difference between V (in the primary studies) and T^2 in the meta-analysis. In the primary study, we treat the observed score for each subject as being the same as the true score for that subject. Therefore, V is the variance of observed effects and also the variance of true effects. By contrast, in the meta-analysis, we distinguish between V (the variance of observed effects) and T^2 (the variance of true effects). The R^2 statistic is based on the variance of true effects. As such, it ensures that (as in the primary study) R^2 can range from 0.00 to 1.00.

Consider the example shown in Figure 12, where we've used Dose as the sole covariate.

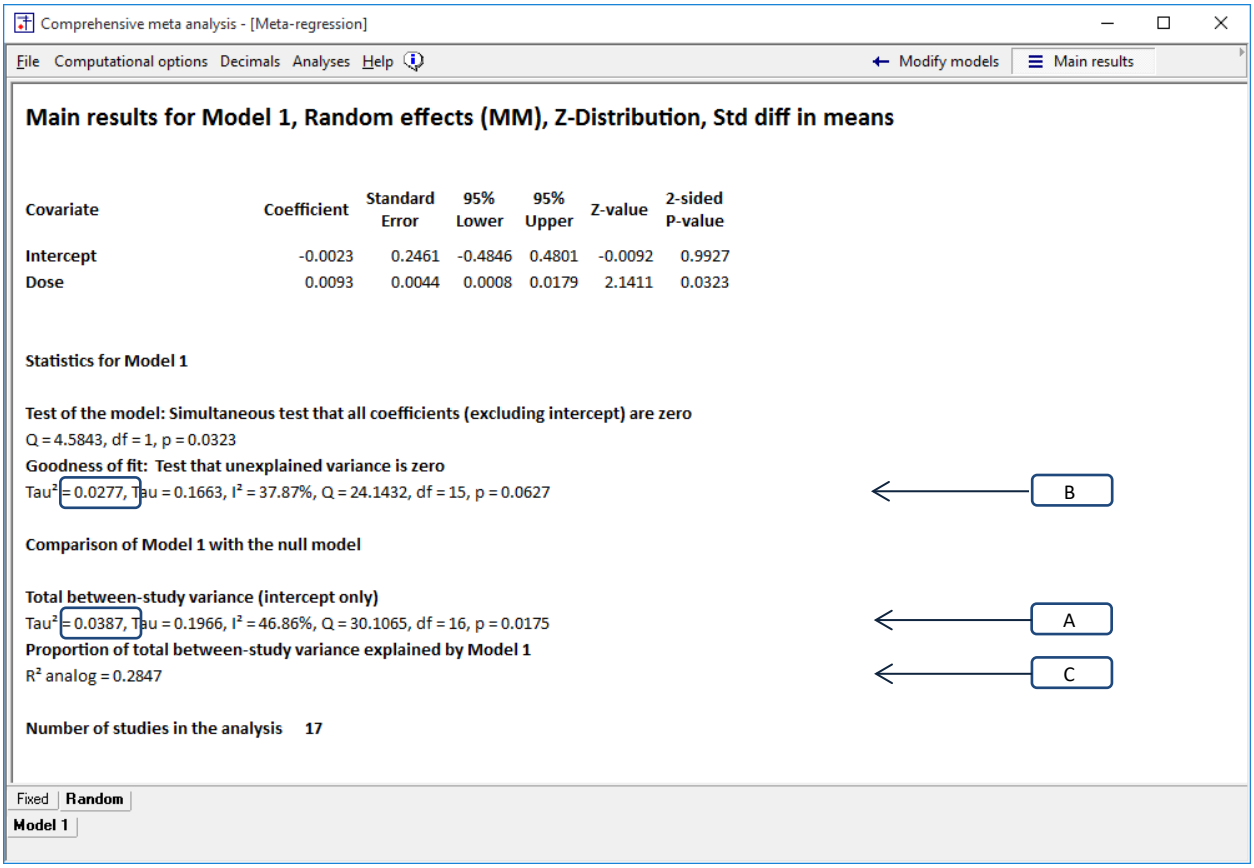


Figure 12 | Main results | Dose | Random-effects

T^2_{Total} is the variance of true effects about the grand mean. To get this value we run the regression with no covariates. When there are no covariates, the predicted value for each study is simply the intercept (which is the grand mean), and 100% of the variance is unexplained. This value [A] is 0.0387.

T^2_{Err} is the variance of true effects about the regression line when we include the covariates. The variance of true effects about the regression line is the unexplained (error) variance. This value [B] is 0.0277.

If we start with the original variance and subtract the variance that remains unexplained by the covariates, the difference is the amount of variance explained by the covariates. Here,

$$T^2_{Explained} = T^2_{Total} - T^2_{Err} = 0.0387 - 0.0277 = 0.0110 \tag{1.8}$$

Then, we compute R^2 using either (1.6)

$$R^2 = \frac{T^2_{Explained}}{T^2_{Total}} = \frac{0.0110}{0.0387} = 0.2847\,, \tag{1.9}$$

or equivalently, using (1.7)

$$R^2 = 1 - \left(\frac{T^2_{Err}}{T^2_{Total}} \right) = 1 - \left(\frac{0.0277}{0.0387} \right) = 0.2847 , \tag{1.10}$$

which is the value reported by the program [C].

The program elaborates on this as shown in Figure 13. To navigate to this screen, click More > R-squared graphic.

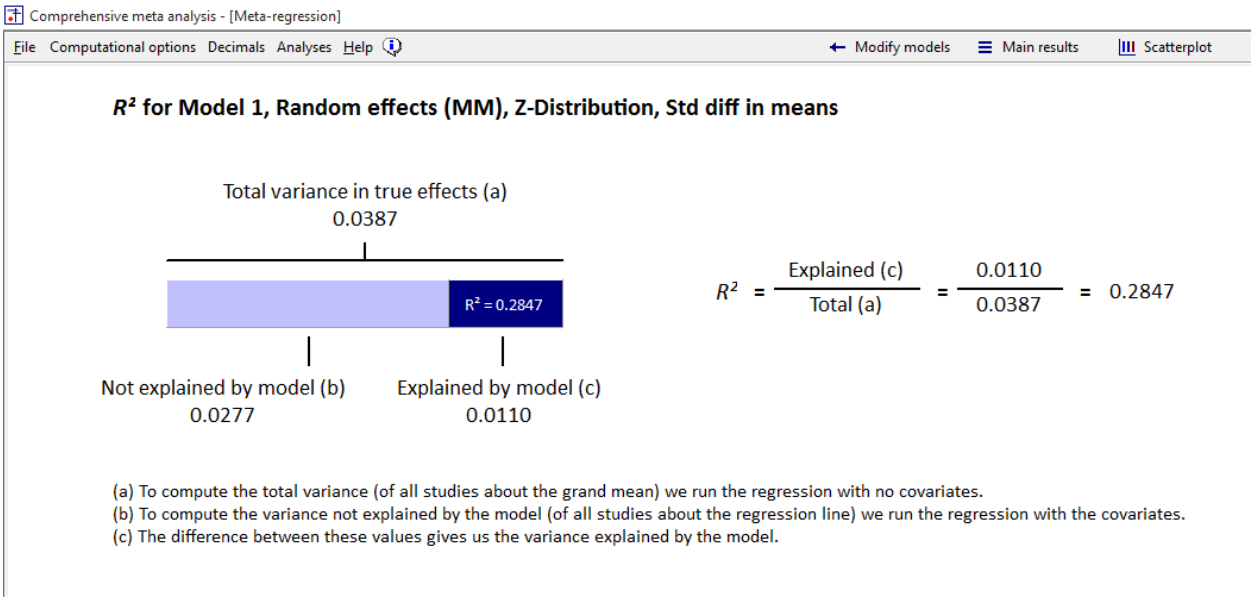


Figure 13 | Display R^2

At the left, the entire bar corresponds to the total variance (0.0387). The light part of the bar reflects the variance not explained by the model (0.0277). The dark part reflects the variance explained by the model (0.0110). Then, R^2 is simply the proportion of the bar that is dark. This is labelled $R^2 = 0.2847$.

Toward the right, the screen shows the computation as

$$R2 = \frac{T^2_{Explained}}{T^2_{Total}} = \frac{0.0110}{0.0387} = 0.2847 . \tag{1.11}$$

To test R^2 for statistical significance we can pose the null that R^2 is zero. This is equivalent to the null that the model explains none of the variance, and so the same test applies. For the model, $Q = 4.5843$, $df = 1$, and $p = 0.0323$ [D]. We conclude that R^2 in the population is probably not zero.

Finally, consider Figure 14, which includes two plots.

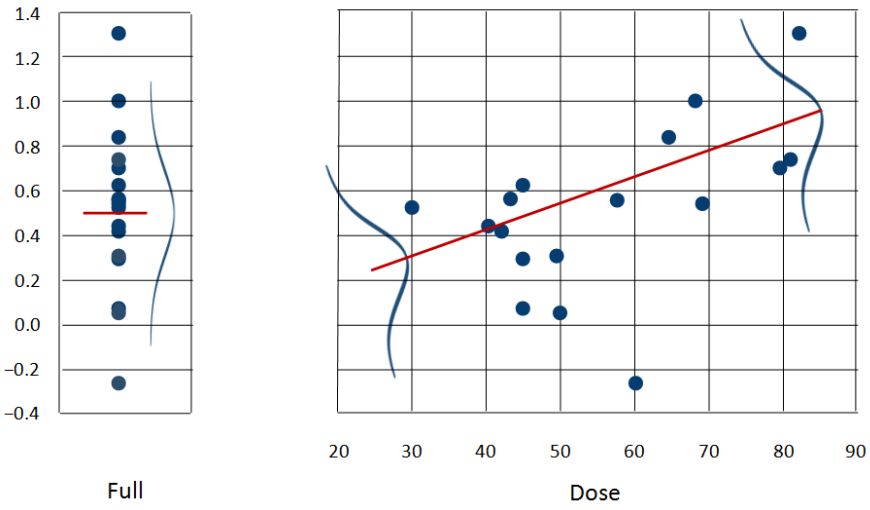


Figure 14 | Dispersion of effects about grand mean vs. dispersion of effects about regression line

At the left we show the dispersion of effects about the grand mean. We've added a normal curve which extends two standard deviations on either side of the mean, and as such is intended to capture some 95% of all true effects. The standard deviation (*T*) is 0.1966, and the curve extends 2*T* (0.39) on either side of the mean.

At the right we show the dispersion of effects about the regression line. We've added a normal curve which extends two standard deviations on either side of the mean, and as such is intended to capture some 95% of all true effects. The standard deviation (*T*) is 0.1663 and the curve extends 2*T* (0.33) on either side of the regression line.

The fact that the curve at right covers a smaller distance than the one at left, illustrates the fact that the covariates improve our ability to predict the effect size. Put simply, the dispersion of effects about the regression line (at right) is less than the dispersion of effects about the grand mean (at left).

While the curve at right is less high than the one at left, the difference between them does not correspond directly to R^2 (or more correctly, to $1 - R^2$). This is because R^2 reflects the ratio of *variances*, while the difference between the curves reflects the ratio of *standard deviations*. The ratio of the variances is

$$1-R^2=\frac{T_{Err}^2}{T_{Total}^2}=\frac{0.0277}{0.0387}=0.7158 \tag{1.12}$$

(which corresponds to $1 - R^2$). By contrast, the ratio of the standard deviations is

$$1-R^2=\frac{T_{Err}}{T_{Total}}=\frac{0.1664}{0.1967}=0.8460 \tag{1.13}$$

(which corresponds to the square root of 0.7158). For this reason, the line at right is 85% as high as the one at left, rather than 72% as high.

As noted, we would expect that some 95% of all true effects are expected to fall within the curve. Sometimes the curves will appear to follow this rule, but sometimes they will not. There are two reasons for this.

First, this statistic refers to 95% of all *true* effects, and not 95% of all *observed* effects. The observed effects tend to be dispersed more widely than true effects (due to sampling error), and so more than 5% of observed effects will fall outside the curve. In the extreme case, if T^2 is zero, then all true effects are expected to fall directly at the mean (at left) or directly on the regression line (at right), and so the curve would have zero width, but the observed effects will vary about these points.

Second, T^2 (and T) are estimates of the variance (and standard deviation) for all populations in the sampling frame, not just the ones in the analysis. The range of effects in a small sample of populations will often differ from the range of effects across a large number of populations.

For these reasons, one should not pay a lot of attention to the precise height of the lines, but only to general concept. To wit, if R^2 is near zero, the curves at right will be approximately the same height as the one at left. As R^2 increases the curves at right will become increasingly narrow. In the extreme case (where R^2 is 1.0) the curves at right will have zero width, suggesting that we can predict the true effect size for each study without any error.

Error in estimating R^2

In a meta-analysis, we estimate the value of T^2 in the sample, and this estimate is imperfect. In approximately half the cases, our estimate will be too low, and in the other half it will be too high (see box). This has implications for the estimation of R^2, as follows.

If we under-estimate the initial variance, and/or over-estimate the final variance, we will tend to under-estimate R^2. Conversely, if we over-estimate of the initial variance and/or under-estimate the final variance, we will tend to over-estimate R^2. The problem will be more acute when we have a small number of studies, and less reliable estimates of T^2.

A related issue is the fact that when T^2 is close to zero and we under-estimate the true value, our estimated value may be zero. If the initial value of T^2 is estimated as zero then R2 will be computed as zero (since there is no variance to explain). Conversely, if the final value of T^2 is estimated as zero then R^2 will be estimated as 1.00 (since it will appear that all the variance has been explained). When R^2 is estimated as 0.00 or as 1.00 we should understand that this is due in part to the role of error in the estimates.

To understand why we under-estimate or over-estimate T^2, we need to understand how T^2 is estimated. We compute the observed dispersion (*Q*), subtract the value of *Q* we would expect to see based on sampling error alone (*df*), and the difference (*Q* minus *df*) is assumed to reflect variance in true effects.

Critically, when we subtract *df*, we are subtracting the *expected* value of the sampling error variance. *On average*, over an infinite number of meta-analyses, the *Q* value due to sampling error will be close to *df*. But in any single meta-analysis, the *Q* value due to sampling error will be less than df or greater than df.

If we happen to have a set of studies where the sampling error was less than the expected value, then (when we subtract *df*) we attribute too much of the observed dispersion to sampling error, and under-estimate the true variance. Conversely, if we happen to have a set of studies where the sampling error exceeds the expected value, then (when we subtract df) we attribute too little of the observed dispersion to sampling error, and over-estimate the true variance.

When we compute R^2 we estimate T^2 twice, and these estimates are independent of each other. This can lead to cases where we underestimate one value while overestimating the other.

BUILDING A MODEL

In meta-analysis, as in a primary study, the basic rules of a regression are as follows. When the regression includes only the intercept, the intercept gives us the mean effect size. When we add covariates to the equation, the coefficient for the covariates tell us how the covariates are related to the effect size. Suppose that we have two variables, X_1 and X_2.

- If we run a regression with X_1 only, the regression will report the relationship between X_1 and the effect size, ignoring any confound with X_2.

- If we run a regression with X_2 only, the regression will report the relationship between X_2 and the effect size, ignoring any confound with X_1.

- If we run a regression with X_1 and X_2, the regression will report the relationship between X_1 and the effect size with X_2 held constant; the relationship between X_2 and the effect size with X_1 held constant; and the relationship between X_1 and X_2 (as a set) and effect size.

To show how this works, we will run a series of analyses, as follows.

1. First, we will review the basic analysis of the ADHD data – the simple analysis, with no regression. We will use this analysis to explain the meaning of the various heterogeneity statistics. This will provide the foundation needed to understand the (same) statistics in the context of regression.

2. Second, we will run a regression that includes the intercept but no covariates. We'll use this to show the correspondence between the statistics reported in a regression and those reported for the simple analysis. When there are no covariates, the statistics are identical, which makes the correspondence more obvious.

3. Third, we will run a regression using Dose as the covariate. This will allow us to assess the impact of Dose, ignoring the potential confound with SUD.

4. Fourth, we will run a regression using SUD as the covariate. This will allow us to assess the impact of SUD, ignoring the potential confound with Dose.

5. Fifth, we will run a regression using both Dose and SUD as covariates. This will allow us to assess the impact of Dose holding SUD constant; the impact of SUD holding Dose constant; and the combined impact of Dose and SUD.

ANALYSIS 1 – A REGRESSION WITH NO COVARIATES

The basic analysis is displayed in Figure 15, with the corresponding statistics displayed in **Error! Reference source not found.** and **Error! Reference source not found.**.

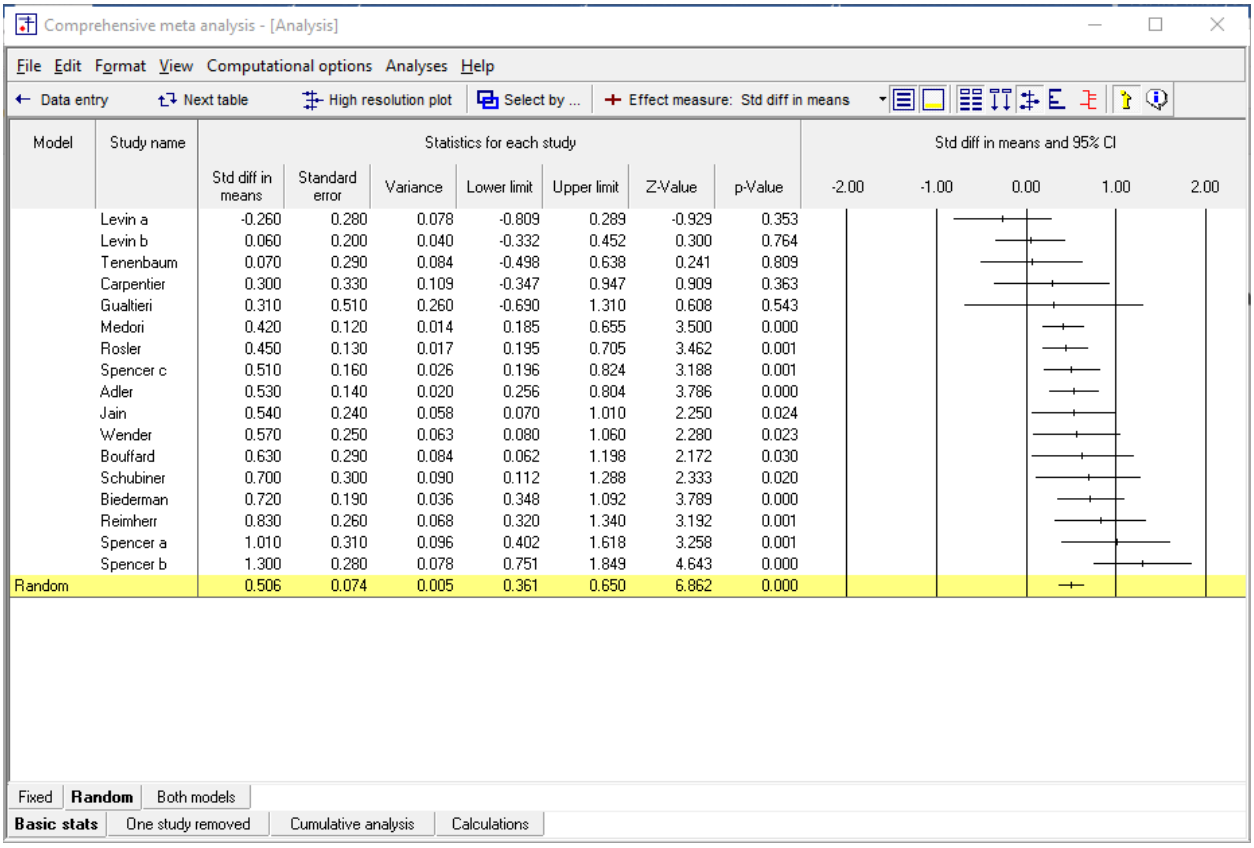
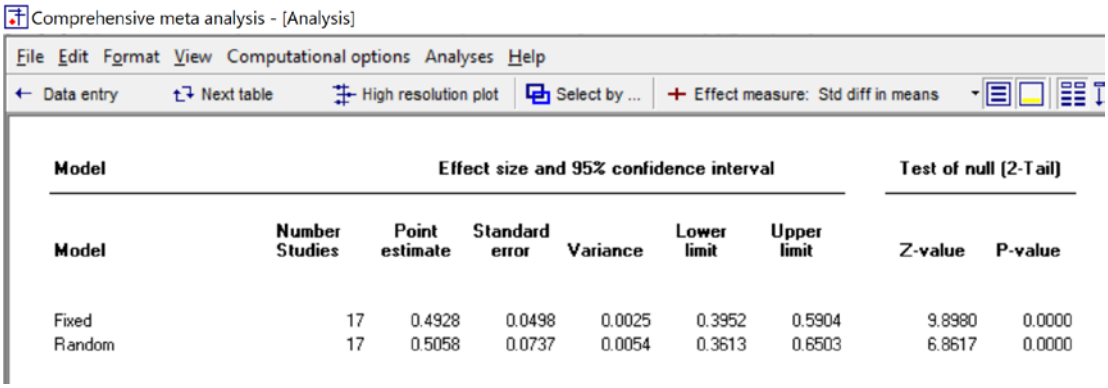


Figure 15



Model		Effect size and 95% confidence interval					Test of null (2-Tail)	
Model	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value
Fixed	17	0.4928	0.0498	0.0025	0.3952	0.5904	9.8980	0.0000
Random	17	0.5058	0.0737	0.0054	0.3613	0.6503	6.8617	0.0000

Figure 16

Heterogeneity				Tau-squared			
Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
30.1065	16.0000	0.0175	46.8553	0.0387	0.0310	0.0010	0.1966

Figure 17

The statistics in Figure 16 address the *mean* effect size. In the sample of seventeen studies, the mean effect size is 0.5058. The studies in our analysis have been sampled from a universe of populations, and we use the sample estimate to generalize to that universe. The confidence interval is 0.3613 to 0.6503, which tells us that the mean effect size in the universe probably falls somewhere in this interval. We might also test the null hypothesis that the mean effect size in the universe is zero. The *Z*-value for this test is 6.8617, with a corresponding p-value of < 0.0001. We reject the null hypothesis, and conclude that the mean effect size is greater than zero.

The statistics in Figure 17 address the issue of heterogeneity – how the effects are dispersed about the mean. To provide context for these statistics, recall that we need to distinguish between the variance of observed effects and the variance of true effects (see chapter ____).

In a meta-analysis, we can pose the null hypothesis that all studies share the same *true* effect size. While the observed effect sizes clearly vary from study to study, the null hypothesis asserts that all studies share the same *true* effect size and that all variation in observed effects is due to sampling error.

The *Q*-value and its degrees of freedom may be used to test this null hypothesis. If the null hypothesis is true, then *Q* will be distributed approximately as chi-squared with *k*−1 degrees of freedom (where *k* is the number of studies). Here, *Q* is 30.1065, the degrees of freedom is 16, and the corresponding *p*-value is 0.0175. For testing this null hypothesis, we typically use a criterion alpha of 0.10. Since the observed *p*-value is less than 0.10 we reject the null, and conclude that the true effect size does vary across studies – methylphenidate has more of an impact in some populations than in others.

This test is sometimes referred to as a "Goodness of fit" test. The "model" would be that the true effect size for all populations is simply the mean effect size. When we reject the null hypothesis, we conclude that the data do not fit the model.

While it's common to test the null hypothesis that the variance in true effects is zero, it's more useful to estimate the extent of this variance. The following statistics quantify specific aspects of the variance.

The I^2 statistic is widely thought to tell us how much the effect size varies across studies, but this is incorrect. Rather, I^2 provides context for the dispersion of observed effects. As outlined above, the dispersion in observed effects tends to exceed the dispersion in true effects. The I^2 statistic tells us what proportion of the variance in observed effects reflects variance in true effects rather than sampling error.

In this example I^2 is 46.8553, which tells us that some 47% of the variance that we see in the forest plot is due to variance in true effects. If we could somehow "shrink" the variance by 53% (that is, by 1 minus 47%), we would see how the true effects vary. While it is technically possible to perform this exercise (see appendix), it's much simpler to compute a prediction interval, which provides the same information in a more direct and intuitive metric. This is discussed below.

T^2 is the variance of true effects, which in this example is 0.0387. This is an estimate of how the true effect size varies across studies in the universe of populations from which our studies were sampled. Since there are no covariates in the analysis, all of this variance is unexplained – we don't know why the effect size varies across studies, only that it does vary. When we move on to meta-regression, we will try to explain some of this variance.

T is the standard deviation of true effects, which in this example is 0.1966. *T* is simply the square root of T^2. This is an estimate of how the true effect size varies across populations, but in linear units rather than squared units. If we assume that our estimates of the mean and *T* are accurate, and that the true effects are normally distributed about the mean, then we would expect most true effects to fall within some two *T* on other side of the mean. Here,

$$\begin{aligned} PI_{LL} &= M - 2T = 0.5058 - 2\left(0.1966\right) = 0.1126 \\ PI_{UL} &= M + 2T = 0.5058 + 2\left(0.1966\right) = 0.8990 \end{aligned} \tag{1.14}$$

This simple formulas is intended to explain the role of *T*. Later, we discuss formulas for computing the prediction interval, taking into account the fact that *M* and *T* are estimated with error.

Summary

We used a simple meta analysis to assess the impact of treatment on cognitive scores.

The summary effect size is 0.5058, which tells us that the mean effect size in our sample is 0.5058. The confidence interval is 0.3613 to 0.6503, which tells us that the mean effect size for the universe of comparable populations probably falls in this interval. We may pose the null hypothesis that the mean effect size is 0.0000, and test this using *Z*. The *Z*-value is 6.8617 and the corresponding p-value is < 0.0001. We reject the null hypothesis, and conclude that the mean effect size is greater than zero.

T, the estimate of the standard deviation of *true* effects, 0.197. T^2, the estimate of the variance of true effects, is 0.039. The *Q*-value, along with its degrees of freedom and p-value, provides a test of the null hypothesis that the parameters τ and τ^2 are zero. Here, *Q* is 30.106 with 16 degrees of freedom and a *p*-value of 0.017. We reject the null hypothesis and conclude that the true effect size varies across studies. Finally, I^2 is 46.855, which tells us that some 47% of the variance in observed effects reflects variance in true effects rather than random error variance.

ANALYSIS 2 – A REGRESSION WITH NO COVARIATES

While we generally use regression to assess the impact of covariates, it's also possible to run a regression with only the intercept, and no covariates.

This adds nothing to the prior analysis. We include it only to show the correspondence between the simple analysis and this regression. Specifically, the statistics in this regression will be identical to those reported in Analysis−1.

The regression results are displayed in Figure 18 and plotted in Figure 19.

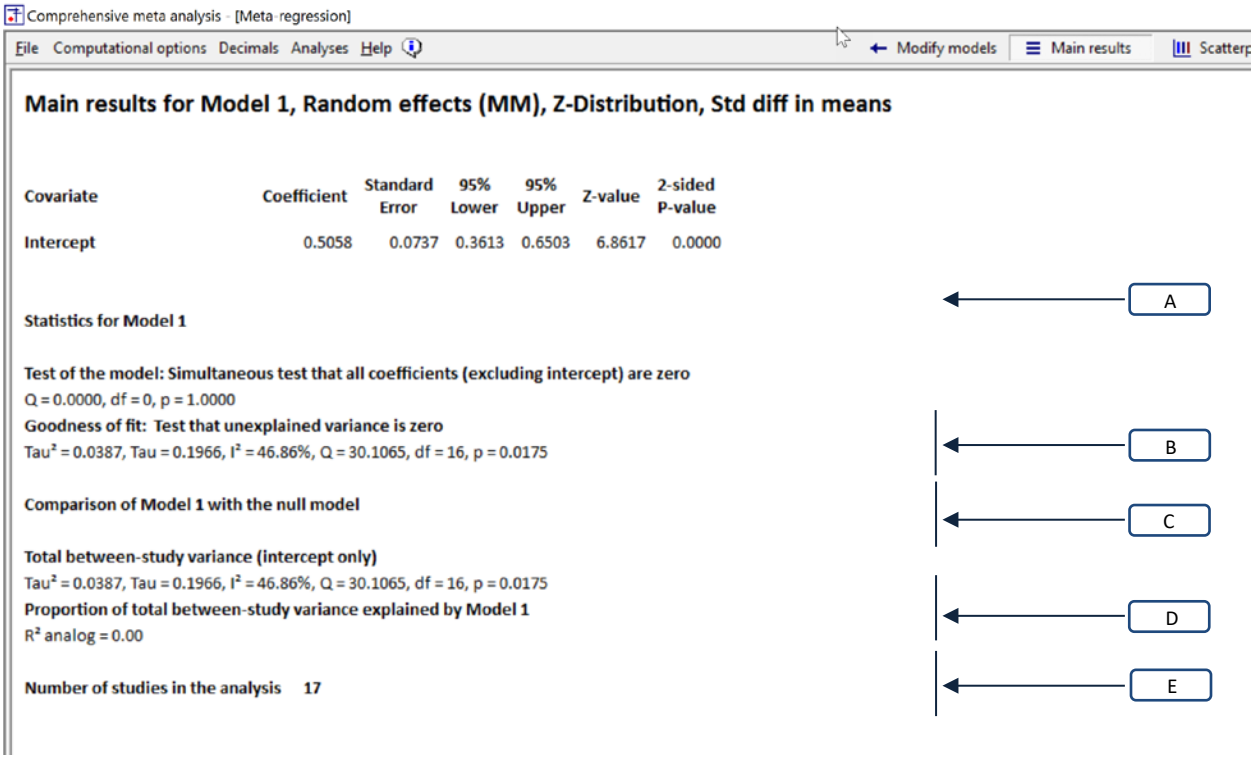


Figure 18

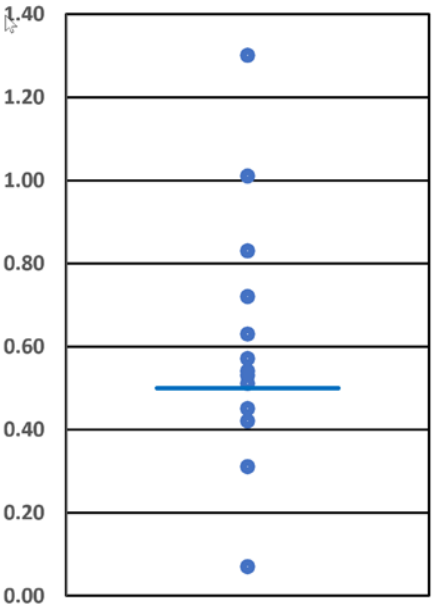


Figure 19

When we have no covariates, the predicted effect size will be the intercept. And, the intercept will be the mean effect size. It follows that the statistics for the intercept in the Analysis-2 will be identical to those for the mean in Analysis-1.

In Figure 18, the intercept [A] is 0.5058, which tells us that the mean effect size in our sample is 0.5058. The confidence interval is 0.3613 to 0.6503, which tells us that the mean effect size for the universe of comparable populations probably falls in this interval. We may pose the null hypothesis that the mean effect size is 0.0000, and test this using *Z*. The *Z*-value is 6.8617 and the corresponding p-value is < 0.0001. We reject the null hypothesis, and conclude that the mean effect size is greater than zero. The statistics on this line correspond exactly to the statistics presented for the mean in Analysis-1 (**Error! Reference source not found.**).

Statistics for Model 1

Test of the model

The section labeled "Test of the model" [B] addresses the question "Is the model is able to explain *any* of the variation in effect size?" The null hypothesis being tested is that none of the covariates are related to effect size.

The test of the model [B] is not relevant as there is no model being tested.

Goodness of fit

This covers sections [C] in Figure 20.

The statistics in this section address the variation of effects about the regression line. The Q-value is the sum of squared deviations about the regression line (on a standardized scale). If all true effects fall directly on the regression line, and all variance is due to sampling error, the expected value of Q is the degrees of freedom (*df*). By working with *Q* and *df* we can estimate the variance of true effects.

Where section B asked if the model is able to explain *any* of the variance in effect size, section C asks if the model is able to explain *all* of the variation in effect size. Concretely, we can pose the null hypothesis that the true effect size for each study falls directly on the regression line, and all variation about the regression line is due entirely to the fact that the observed effects (which are the ones on the plot) differ from the true effects. If the null hypothesis is true, Q will be distributed approximately as chi-squared. When we compare Q to a chi-squared distribution with the appropriate df, the corresponding p-value reflects the test of the null.

This section is called "Goodness of fit" because the null hypothesis is that model is able to explain all the variance – the data "fits" the model.

In Figure 19 the regression line is horizontal. This follows from the fact that the predicted value for each study is simply the intercept. The effects vary about the regression line, and this variance is addressed by section [C] in Figure 18.

T is the standard deviation of *true* effects, which in this example is 0.197. T^2 is the variance of true effects, which is simply the square of the standard deviation. In this example, T^2 is 0.039. The *Q*-value, along with its degrees of freedom and p-value, provides a test of the null hypothesis that the variance of true effects is zero. In our example, *Q* is 30.106 with 16 degrees of freedom and a *p*-value of 0.017.

Finally, I^2 is reported as 46.855%. As always, I^2 does not tell us how much the effects vary. Rather, it provides some context for the variance of observed effects. While the plot shows some variation in observed effects about the regression lines, only about 47% of this reflects variation in true effects. If we could somehow get rid of the sampling error (if we could plot the true effect size for each study) the effects would tend to fall closer to the regression line than the ones we see in the plot.

The previous paragraph is an abbreviated version of the explanations given for Analysis−1. The key point is that the statistics here have the same meaning as those in Analysis-1 earlier and (it follows) the same values. In subsequent analyses, we will add covariates to the model, and these numbers will change.

Comparison of Model 1 with the null model

When we perform a regression analysis in a primary study we often report a statistic called R^2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T^2_0 is the variance of true effects about the mean, and T^2_1 is the variance of true effects about the

regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

Total between-study variance (intercept only)

Section [D] gives us the same statistics for a model based only on the intercept. In this case, where the main model includes no covariates, section [D] is the same as section [C].

Proportion of variance explained by Model-1

Section [E] reports the proportion of variance explained by the model (i.e., by the covariates). Since there are no covariates in the model, this section is not applicable, and the proportion explained is shown as zero.

Summary

We used meta-regression with no covariates to assess the impact of treatment on cognitive scores.

The intercept is 0.5058, which tells us that the mean effect size in our sample is 0.5058. The confidence interval is 0.3613 to 0.6503, which tells us that the mean effect size for the universe of comparable populations probably falls in this interval. We may pose the null hypothesis that the mean effect size is 0.0000, and test this using *Z*. The *Z*-value is 6.8617 and the corresponding p-value is < 0.0001. We reject the null hypothesis, and conclude that the mean effect size is greater than zero.

T, the estimate of the standard deviation of *true* effects, 0.197. T^2, the estimate of the variance of true effects, is 0.039. The *Q*-value, along with its degrees of freedom and p-value, provides a test of the null hypothesis that the parameters τ and τ^2 are zero. Here, *Q* is 30.106 with 16 degrees of freedom and a *p*-value of 0.017. We reject the null hypothesis and conclude that the true effect size varies across studies. Finally, I^2 is 46.855, which tells us that some 47% of the variance in observed effects reflects variance in true effects rather than random error variance.

ANALYSIS 3 – DOSE AS THE COVARIATE

As is true in a primary study, when we include covariates, the goal of the regression is no longer to predict the mean effect, but rather to determine how the effect size is related to the covariates. Therefore, our focus shifts away from the intercept, and toward the statistical model.

First, let's consider an analysis with Dose as the only covariate. This analysis will assess the relationship between Dose and effect size, ignoring the potential impact of any confounds. The results are displayed in Figure 20 and plotted in Figure 21.

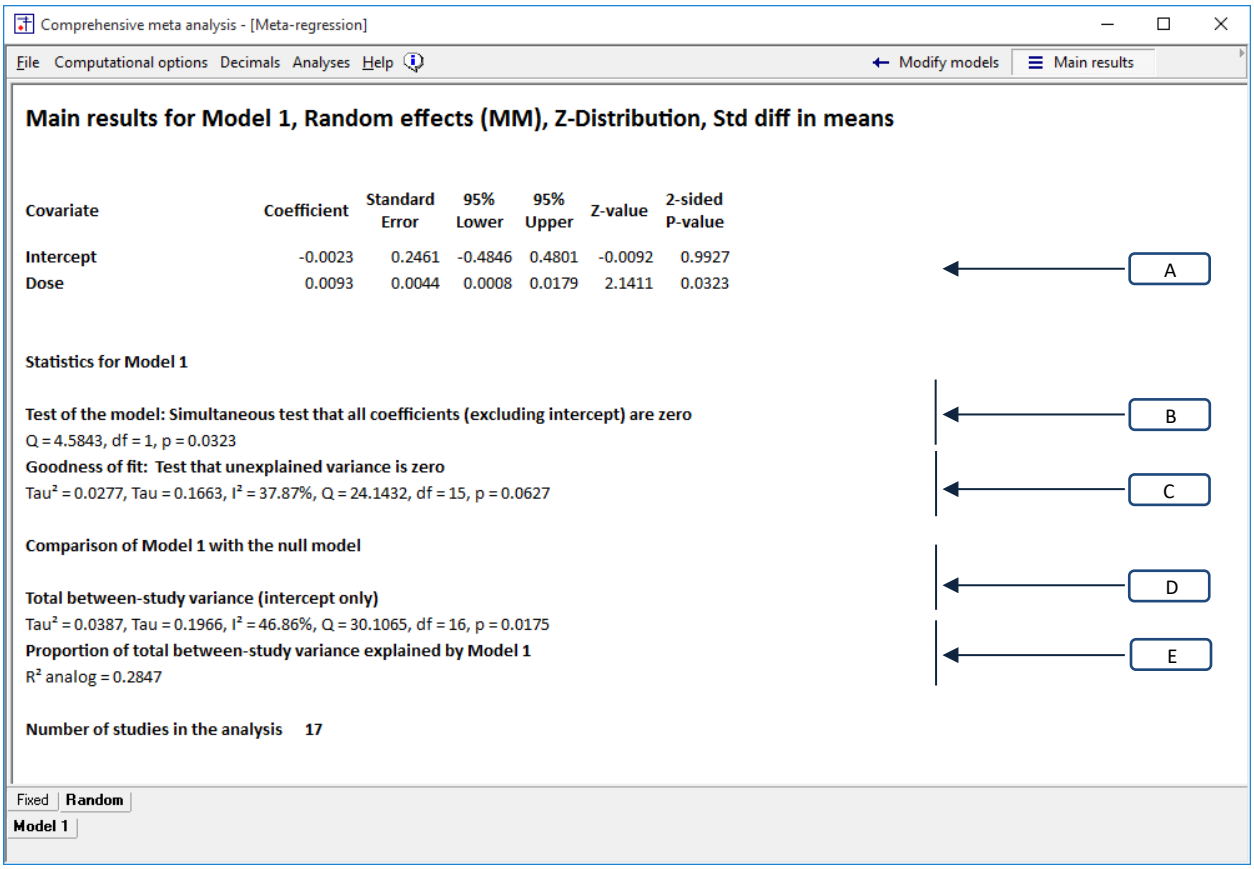


Figure 20

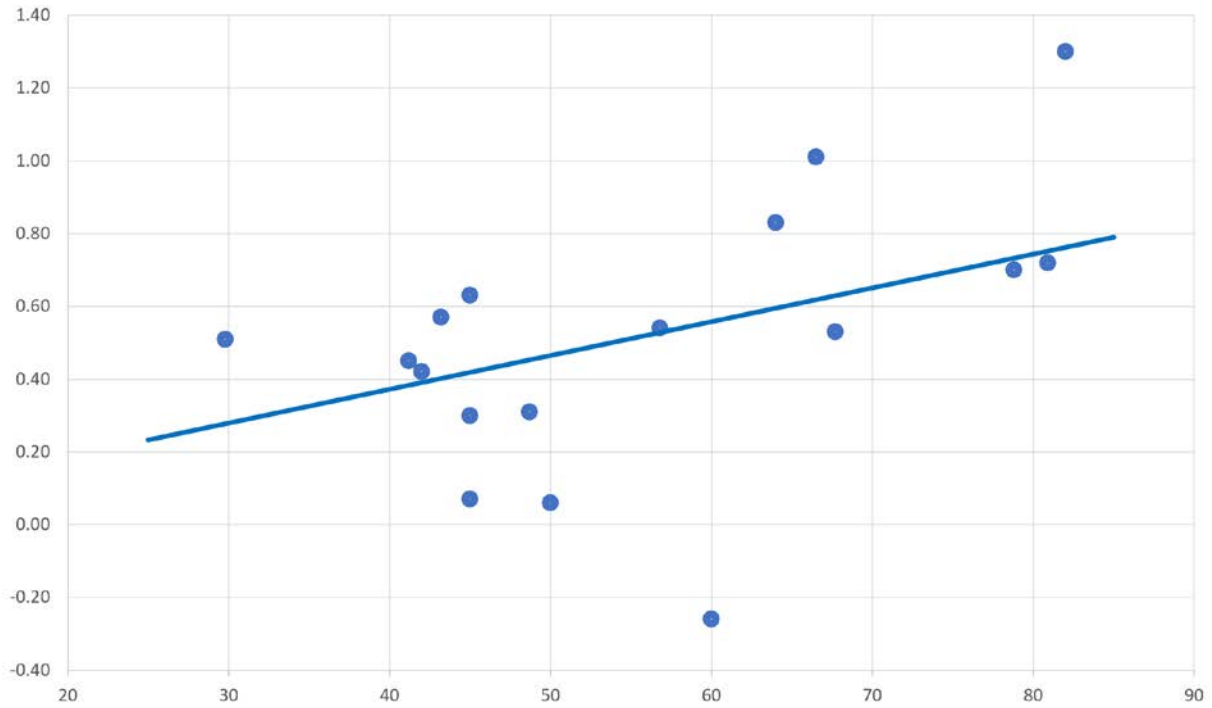


Figure 21

The covariates

Here, we address section [A] in Figure 20.

The line for Dose displays the relationship between Dose and effect size, when there are no other covariates in the model. The coefficient is 0.0093. The fact that the coefficient is positive tells us that a higher dose is associated with a higher effect size. Specifically, a one unit increase in dose corresponds to an increase of 0.0093 in effect size. To make this more intuitive, we could multiply the coefficient by 50, and say that a 50-point increase in dose corresponds to an increase of around 0.47 in the *d*-value. Equivalently, in Figure 21, as the Dose increases by fifty units (from 30 to 80), the predicted effect increases by 47 points (from 0.27 to 0.74).

The coefficient reported here is an estimate of the parameter (the population value), and the other statistics on this line speak to the precision of this estimate. The standard error is 0.0044. If we assume that the true coefficient usually (in 95% of cases) falls within 1.96 standard error of the estimate, then it probably falls in the range of 0.0008 to 0.0179, which is displayed as the 95% confidence interval. At the lower end, a one unit increase in dose would be associated with an increase of 0.0008 in effect size (and a 50-point increase in dose would be associated with an increase of around 0.04 in the *d*-value). At the upper end, a one unit increase in dose would be associated with an increase of 0.0179 in effect size (and a 50-point increase in dose would be associated with an increase of around 0.90 in the d-value).

Finally, we can test the null hypothesis that the true value of the coefficient is zero (that there is no relationship between Dose and effect size). The 95% confidence interval (0.0008 to 0.0179) does not include zero. Similarly, the *Z*-value (computed as the coefficient divided by its standard error) is 2.1411, and the corresponding p-value is 0.0323 (which is less than the criterion alpha of 0.05).

In either case (based on either the confidence interval or the p-value), we reject the null hypothesis that the true value of the coefficient is zero, and conclude that there probably is a positive relationship between dose and effect size.

Statistics for Model 1

Test of the model

The section labeled "Test of the model" [B] addresses the question "Is the model is able to explain *any* of the variation in effect size?" The null hypothesis being tested is that none of the covariates are related to effect size.

The *Q*-value is 4.5843, with one degree of freedom (corresponding to the one covariate) and a *p*-value of 0.0323. Since the *p*-value is less than the criterion alpha of 0.05, we reject the null hypothesis and conclude that at least one of the covariates is related to the effect size.

In this example (since there is only one covariate) the test of the model [B] is identical to the test of that covariate [A]. The test presented in [B] is based on *Q*, which is a squared metric, whereas the one presented in [A] is based on *Z*, which is a linear metric. Therefore, *Z* is reported as 2.1411 while *Q* is reported as 4.5843, which is simply *Z* squared. In any event, the *p*-value is reported as 0.0323 in both cases.

Goodness of fit

This covers sections [C] in Figure 20.

The statistics in this section address the question "Is able to explain *all* of the variation in effect size?" The null hypothesis is that the true effect size for each study falls directly on the regression line, and all variation about the regression line is due entirely to the fact that the observed effects (which are the ones on the plot) differ from the true effects.

The statistics presented here have the same interpretation as those presented in Analysis−2. In both cases, the statistics describe the dispersion of effects about the regression line. In Analysis-2, the regression line was horizontal – the predicted effect size for all studies was the same. In the current analysis, the regression line is based on Dose – the predicted effect size is higher for studies with a higher Dose. But in both cases, the statistics describe the dispersion of true effects about the regression line, and test the null hypothesis that all true effects fall directly on the regression line.

The standard deviation of true effects, *T*, is 0.1663. If we assume that most true effects will fall within roughly two standard deviations of the regression line, then most will fall within 2*T* (0.33) of the regression line. We can get a sense of this by looking at the plot, and visualizing a normal curve that extends two standard deviations on either side of the regression line. For a study where the Dose is 30 units, the true effect size will usually fall in the range of −0.05 to +0.61. For a study where the Dose is 80 units, the true effect size will usually fall in the range of 0.41 to +1.07.

The variance of true effects, T^2, is 0.0277. This is simply the standard deviation, squared. This is a less intuitive metric than *T* but it has some useful statistical properties.

The *Q* statistic may provide a test of the null hypothesis that the true effect size for all studies is the predicted value. This could also be framed as "The true effect size for all studies lies precisely on the regression line."

The *Q*-value is 4.5843 with 15 degrees of freedom and a p-value of 0.0627. When we test for heterogeneity we typically use a criterion alpha of 0.10, and on this basis we would reject the null. That is, the dispersion of effects about the regression line exceeds the amount we would expect to see based on sampling error alone. Another way to say this, is that the actual value of of τ (and τ 2) is probably not zero.

Finally, I^2 is reported as 37.87%. As always, I^2 does not tell us how much the effects vary. Rather, it provides some context for the variance of observed effects. While the plot shows some variation in observed effects about the regression lines, only about 38% of this reflects variation in true effects. If we could somehow get rid of the sampling error (if we could plot the true effect size for each study) the effects would tend to fall closer to the regression line than the ones we see in the plot.

Comparison of Model 1 with the null model

When we perform a regression analysis in a primary study we often report a statistic called R^2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T^2_0 is the variance of true effects about the mean, and T^2_1 is the variance of true effects about the regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

Total between-study variance (intercept only)

Section [D] reports that with no covariates in the model, the variance of true effects is 0.0387. This is the same value reported in Analysis-1 and Analysis-2, and it will serve as the baseline value for computing R^2.

Proportion of variance explained by Model-1

In section [D] we saw that the variance of true effects about the mean was 0.0377. And in section [C] we saw that the variance of true effects about the regression line based on Dose was 0.0277. The difference

$$T^2_{Explained} = T^2_{Total} - T^2_{Residual} = 0.0387 - 0.0277 = 0.0110$$

is the variance in true effects explained by Dose. We can then estimate the proportion of variance explained by Dose using

$$R^2 = \frac{T^2_{Explained}}{T^2_{Total}} = \frac{0.0110}{0.0387} = 0.28$$

Summary

We used meta-regression to examine the linear relationship between Dose and effect size.

The coefficient for Dose is 0.0093, which means that for every one-unit increase in dose, the effect size increases by 0.0093. As Dose increases from 30 units to 80 units, the expected effect size increases from 0.28 to 0.74. The 95% confidence interval for the coefficient extends from 0.0008 to 0.0179. The *Z*-value for a test of the null (that the true coefficient is zero) is 2.1411, and the corresponding *p*-value is 0.0323. [A]

A test of the model yields a *Q*-value of 4.5843 with 1 degree of freedom and corresponding *p*-value of 0.0323. We conclude that the model is able to explain at least some of the variance in effect size. [B]

The variance of true effects about the regression line (T^2) is 0.0277, and the standard deviation of true effects about the regression line (T) is 0.1663. The I^2 statistic is 37.87%, which tells us that some 38% of the observed variance about the regression line reflects variation in true effects rather than sampling error. The test for heterogeneity yields a *Q*-value of 24.1432 with 15 degrees of freedom and a corresponding *p*-value of 0.0627. We conclude that the model does not fully explain the variation in effects. [C]

The R^2 analog is 0.28, which means that the model is able to explain some 28% of the variance in true effects. [E]

The relationship between Dose and effect size is observational, not causal. It's possible that the higher effects in some studies were caused by dose, but it's also possible that they were caused (in whole or in part) by factors that were confounded with dose.

ANALYSIS 4 – SUD AS THE COVARIATE

Next, let's consider an analysis with SUD as the only covariate, as displayed in Figure 20. This analysis will assess the relationship between SUD and effect size, ignoring the potential impact of any confounds.

The results are displayed in Figure 22 and plotted in Figure 23. Since SUD is a categorical variable, it's represented in the analysis by a dummy variable called SUD|Y. This variable is coded 0 (SUD excluded) or 1 (SUD included). (See chapter ____ for a discussion of categorical variables).

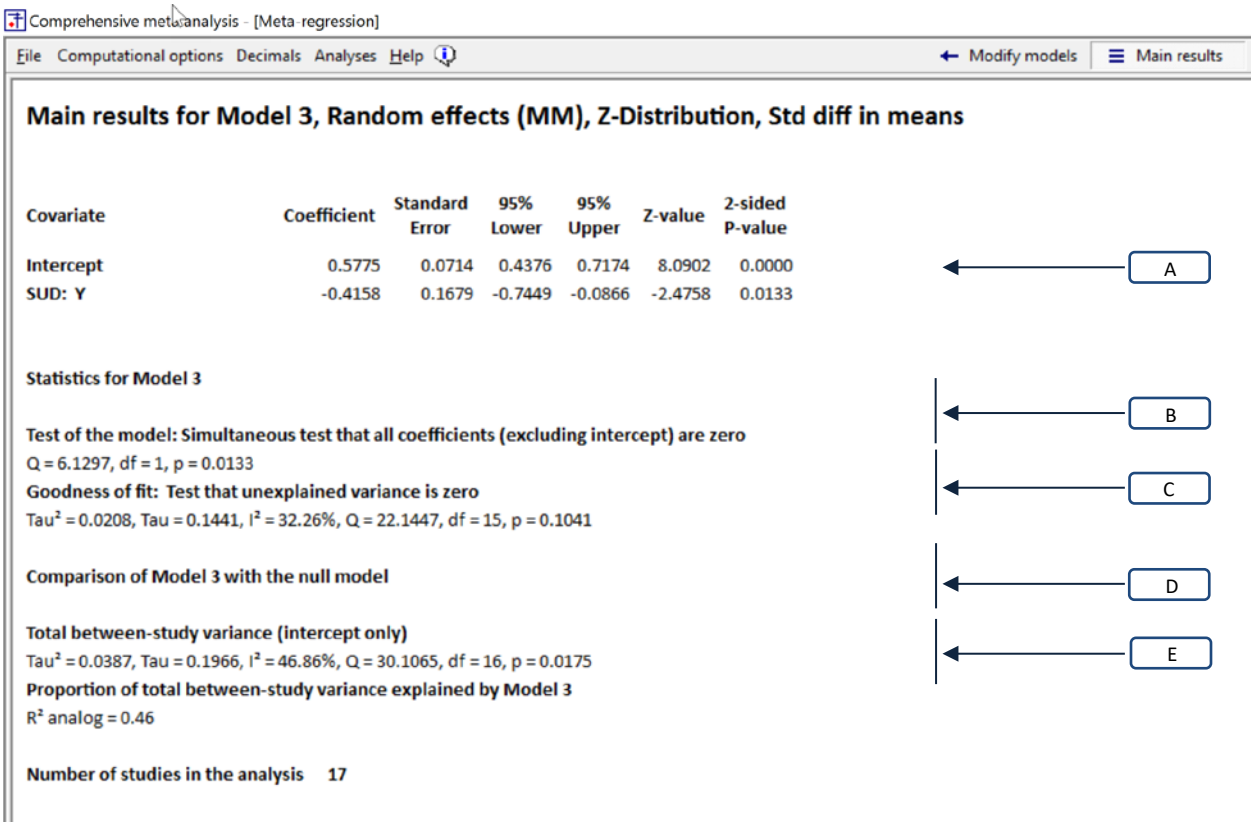


Figure 22

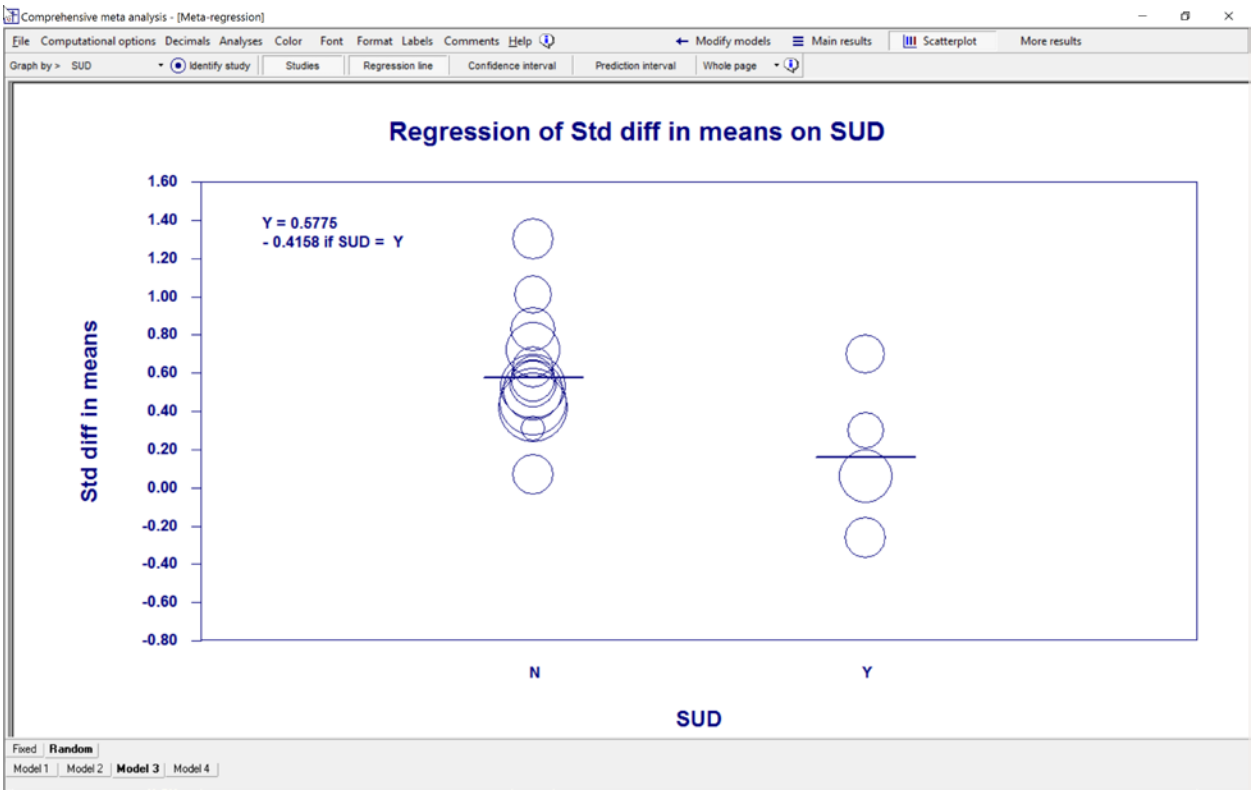


Figure 23

The covariates

Here, we address section [A] in Figure 22

The line for SUD| Y displays the relationship between SUD and effect size, when there are no other covariates in the model.

In chapter _____ we discuss how to interpret coefficient for categorical variables, but for now we can simply say that the coefficient gives us the mean difference in effect size for studies that enrolled SUD patients as compared with those that excluded them. The coefficient −0.4158, which tells us that the mean effect size for studies that enrolled these patients is 0.4158 lower than for studies that included them.

The coefficient reported here is an estimate of the parameter (the population value), and the other statistics on this line speak to the precision of this estimate. The standard error is 0.1679. If we assume that the true coefficient usually (in 95% of cases) falls within 1.96 standard error of the estimate, then it probably falls in the range of −0.7449 to −0.0866, which is displayed as the 95% confidence interval. The mean for the SUD studies could be as little as 0.09 points lower than the non-SUD studies, or as much as 0.74 points lower.

Finally, we can test the null hypothesis that the true value of the coefficient is zero (that there is no relationship between Dose and effect size). The 95% confidence interval (−0.7449 to −0.0866) does not

include zero. Similarly, the Z-value (computed as the coefficient divided by its standard error) is −2.4758, and the corresponding p-value is 0.0133 (which is less than the criterion alpha of 0.05).

In either case (based on either the confidence interval or the p-value), we reject the null hypothesis that the true value of the coefficient is zero, and conclude that the mean score for SUD studies is probably lower than that for non-SUD studies.

Statistics for Model 1

Test of the model

The section labeled "Test of the model" [B] addresses the question "Is the model is able to explain *any* of the variation in effect size?" The null hypothesis being tested is that none of the covariates are related to effect size.

The *Q* value is 6.1297, with one degree of freedom (corresponding to the one covariate) and a *p*-value of 0.0133. Since the p-value is less than the criterion alpha of 0.05, we reject the null hypothesis and conclude that at least one of the covariates is related to the effect size.

In this example (since there is only one covariate) the test of the model [B] is identical to the test of that covariate [A]. The test presented in [B] is based on *Q* which is a squared metric, whereas the one presented in [A] is based on *Z,* which is a linear metric. Therefore, *Z* is reported as −2.4758 while *Q* is reported as 6.1297, which is simply *Z* squared. In any event, the *p*-value is reported as 0.0133 in both cases.

Goodness of fit

This covers sections [C] in Figure 20.

The statistics in this section address the question "Is able to explain *all* of the variation in effect size?" The null hypothesis is that the true effect size for each study falls directly on the regression line, and all variation about the regression line is due entirely to the fact that the observed effects (which are the ones on the plot) differ from the true effects.

The statistics presented here have the same interpretation as those presented in Analysis−2. In both cases, the statistics describe the dispersion of effects about the regression line. In Analysis-2, the regression line yielded the same predicted effect size for all studies. In the current analysis, the regression line is based on SUD, and there is one predicted effect for SUD studies and a different predicted effect for non-SUD studies. But in both cases, the statistics describe the dispersion of true effects about the regression line, and test the null hypothesis that all true effects fall directly on the regression line.

The standard deviation of true effects about the subgroup means, T, is 0.1441. If we assume that most true effects will fall within 1.96 standard deviations of the regression line, then most will fall within 0.28 points of the subgroup means. We can get a sense of this by looking at the plot, and visualizing a normal curve that extends two standard deviations on either side of the subgroup means. For a non-SUD study

the mean effect is around 0.58, the true effect size will usually fall in the range of 0.30 to 0.88. For an SUD study the mean is around _____, the true effect size will usually fall in the range of ____ to ____.

The variance of true effects, T2, is 0.0208. This is simply the standard deviation, squared. This is a less intuitive metric than T but it has some useful statistical properties.

The *Q* statistic may provide a test of the null hypothesis that the true effect size for all studies is the predicted value. This could also be framed as "The true effect size for all studies lies precisely on the regression line." The Q-value is 22.1447 with 15 degrees of freedom and a p-value of 0.1041. When we test for heterogeneity we typically use a criterion alpha of 0.10. We will treat this as meeting the criterion and reject the null hypothesis (others would disagree). We conclude that the dispersion of effects about the regression line exceeds the amount we would expect to see based on sampling error alone. Another way to say this, is that the actual value of of τ (and τ 2) is probably not zero.

Finally, I^2 is reported as 32.26%. As always, I^2 does not tell us how much the effects vary. Rather, it provides some context for the variance of observed effects. While the plot shows some variation in observed effects about the regression lines, only about 32% of this reflects variation on true effects. If we could somehow get rid of the sampling error (if we could plot the *true* effect size for each study) the effects would tend to fall closer to the regression line than the ones we see in the plot.

Comparison of Model 1 with the null model

When we perform a regression analysis is a primary study we often report a statistic called R2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T^2_0 is the variance of true effects about the mean, and T^2_1 is the variance of true effects about the regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

In section [D] we saw that with no covariates in the model, the variance of true effects is 0.0387. This is the same value reported in Analysis-1 and Analysis-2, and it will serve as the baseline value for computing R2. In section [C] we saw that the variance of true effects about the regression line based on SUD was 0.0208. The difference,

$$T^2_{Explained} = T^2_{Total} - T^2_{Residual} = 0.0387 - 0.0208 = 0.0179,$$

is the variance in true effects explained by SUD. We can then estimate the proportion of variance explained by SUD using

$$R^2 = \frac{T^2_{Explained}}{T^2_{Total}} = \frac{0.0179}{0.0387} = 0.46$$

Summary

We used meta-regression to examine the relationship between SUD and effect size.

The coefficient for SUD| Yes is −0.4185. The mean effect size in studies that enrolled SUD patients was 0.4185 lower than for studies which excluded these patients. The mean effect size for studies that excluded these patients is 0.5775, while the mean effect size in studies that included these patients is 0.1590. [A]

The 95% confidence interval for the coefficient extends from −0.7449 to −0.0866, which tells us that the difference between groups probably falls in the range. The *Z*-value for a test of the null (that the true coefficient is zero) is −2.4758, and the corresponding *p*-value is 0.0133. [A]

A test of the model yields a *Q*-value of 6.1297 with 1 degree of freedom and corresponding *p*-value of 0.0133. We conclude that the model is able to explain at least some of the variance in effect size. [B]

The variance of true effects about the subgroup means (T^2) is 0.0208, and the standard deviation of true effects about the subgroup means (*T*) is 0.1441. The I^2 statistic is 32.26%, which tells us that some 32% of the observed variance about the subgroup means reflects variation in true effects rather than sampling error. The test for heterogeneity yields a *Q*-value of 22.1447 with 15 degrees of freedom and a corresponding *p*-value of 0.1041. This is close to the criterion p-value of 0.10, so we conclude that the model does not fully explain the variation in effects. [C]

The R^2 analog is 0.46, which means that the model is able to explain some 46% of the variance in true effects. [E]

The relationship between SUD and effect size is observational, not causal. The lower effect size in the SUD studies could be due to factors that were confounded with SUD. This is a particular concern since the SUD subgroup includes only four studies.

ANALYSIS 5 – DOSE AND SUD AS THE COVARIATES

In Analysis-3 we included only Dose, and this allowed us to assess the impact of Dose while ignoring the potential confound with SUD. In Analysis-4 we included only SUD, and this allowed us to assess the impact of SUD while ignoring the potential confound with Dose. Now, we will run an analysis with both Dose and SUD as covariates. This will allow us to assess the unique impact of each covariate, as well as the combined impact of the two.

The results are shown in Figure 24 and plotted in Figure 25.

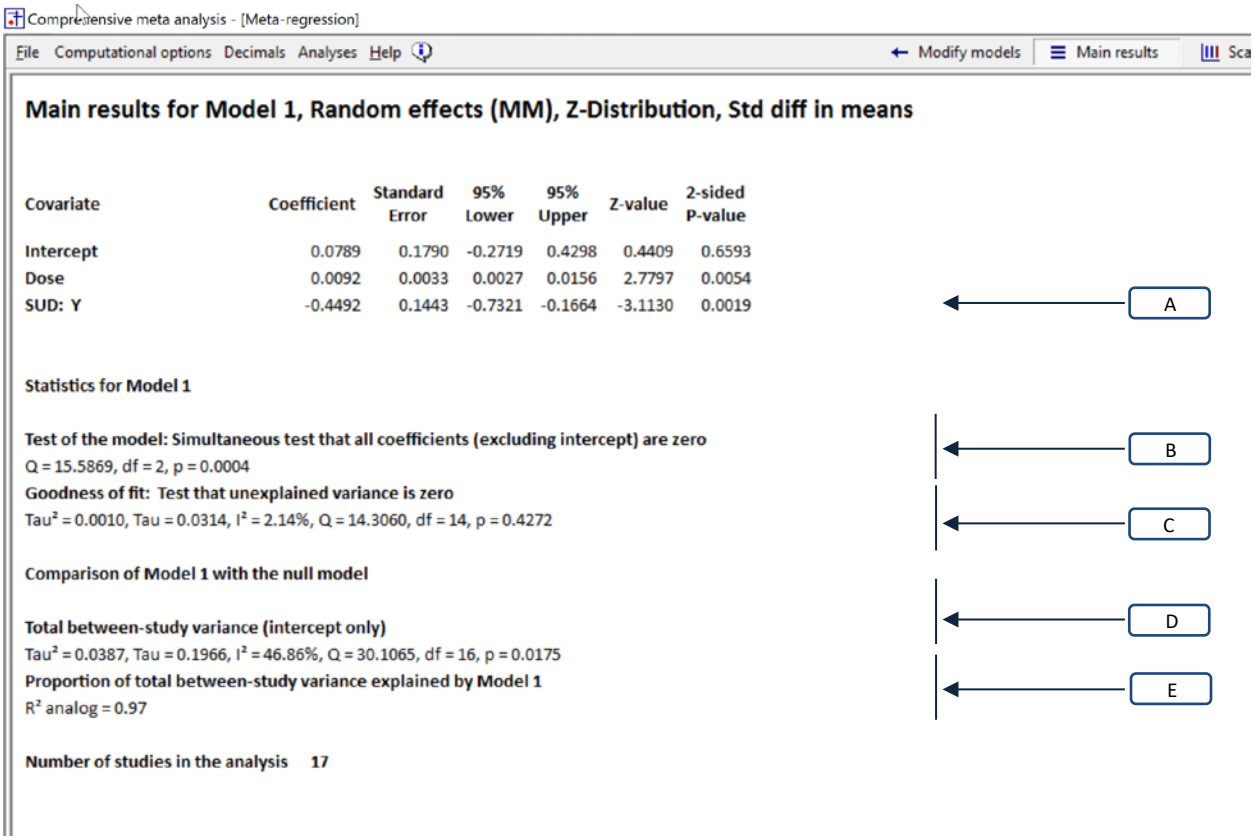


Figure 24

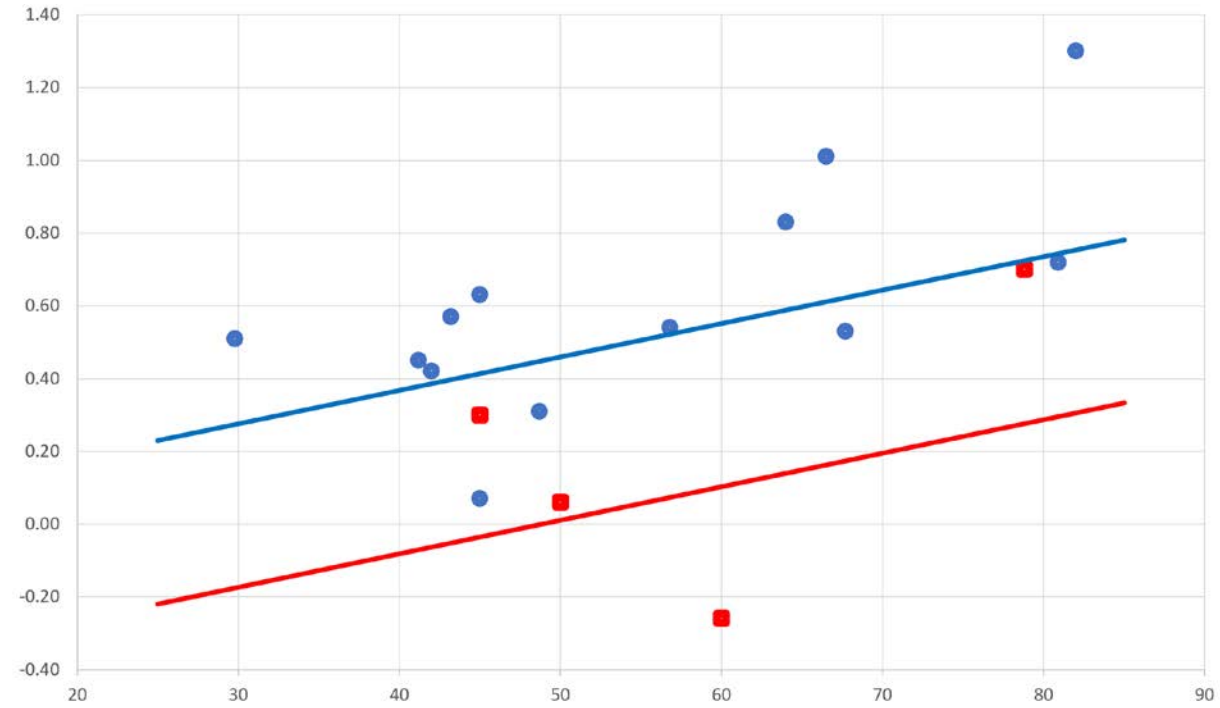


Figure 25

The covariates

Here, we address section [A] in Figure 24.

Relationship between Dose and effect size when SUD is held constant

The line for Dose displays the relationship between Dose and effect size, controlling for SUD. The coefficient for Dose is 0.0092. The fact that the coefficient is positive tells us that a higher dose is associated with a higher effect size. Specifically, a one unit increase in dose corresponds to an increase of 0.0092 in effect size. To make this more intuitive, we could multiply the coefficient by 50, and say that a 50-point increase in dose corresponds to an increase of around 0.46 in the d-value.

The coefficient reported here is an estimate of the parameter (the population value), and the other statistics on this line speak to the precision of this estimate. The standard error is 0.0044. If we assume that the true coefficient usually (in 95% of cases) falls within 1.96 standard error of the estimate, then it probably falls in the range of 0.0027 to 0.0156, which is displayed as the 95% confidence interval. At the lower end, a one unit increase in dose would be associated with an increase of 0.1350 in the d-value (and a 50-point increase in dose would be associated with an increase of around 4 points). At the upper end, a one unit increase in dose would be associated with an increase of 0.0156 in the *d*-value (and a 50-point increase in dose would be associated with an increase of around 78 points).

Finally, we can test the null hypothesis that the true value of the coefficient is zero (that there is no relationship between Dose and effect size when SUD is held constant). The 95% confidence interval (0.0027 to 0.0156) does not include zero. Similarly, the *Z*-value (computed as the coefficient divided by its standard error) is 2.7797, and the corresponding p-value is 0.0054 (which is less than the criterion alpha of 0.05). In either case, we reject the null hypothesis that the true value of the coefficient is zero,

and conclude that there probably is a positive relationship between dose and effect size when SUD is held constant.

Relationship between SUD and effect size when Dose is held constant

The line for SUD| Y displays the relationship between SUD and effect size, when Dose is held constant. In chapter _____ we discuss how to interpret coefficient for categorical variables, but for now we can simply say that the coefficient gives us the mean difference in effect size for studies that enrolled SUD patients as compared with those that excluded them, assuming the same value of Dose. The coefficient −0.4492, which tells us that the mean effect size for studies that enrolled these patients is some 45 points lower than for studies that included them.

THE EFFECT SIZES DIDN'T CHANGE MUCH. THE SE GOT SMALLER

The coefficient reported here is an estimate of the parameter (the population value), and the other statistics on this line speak to the precision of this estimate. The standard error is 0.1443. If we assume that the true coefficient usually (in 95% of cases) falls within 1.96 standard error of the estimate, then it probably falls in the range of −0.7321 to −0.1664, which is displayed as the 95% confidence interval. For a given dose, the mean effect size for the SUD studies could be as little as 0.16 points lower than the non-SUD studies, or as much as 0.73 points lower.

Finally, we can test the null hypothesis that the true value of the coefficient is zero (that there is no relationship between SUD and effect size) when Dose is held constant. The 95% confidence interval (−0.7321 to −0.1664) does not include zero. Similarly, the *Z*-value (computed as the coefficient divided by its standard error) is −3.1130 and the corresponding *p*-value is 0.0019 (which is less than the criterion alpha of 0.05). In either case, we reject the null hypothesis that the true value of the coefficient is zero, and conclude that (with Dose held constant) the mean score for SUD studies is probably lower than that for non-SUD studies.

Statistics for Model 1

Test of the model

The section labeled "Test of the model" [B] addresses the question "Is the model is able to explain *any* of the variation in effect size?" The null hypothesis being tested is that none of the covariates are related to effect size.

In Analysis-3 and Analysis-4 there was only one covariate, and therefore sections [A] and [B] were testing the same model. That's not the case now. In the current analysis each line in section [A] addressed the *unique* impact of one covariate. By contrast, section [B] addresses the *combined* impact of the two covariates.

The *Q* value is 15.5869, with two degrees of freedom (corresponding to the two covariates) and a p-value of 0.0004. Since the p-value is less than the criterion alpha of 0.05, we reject the null hypothesis and conclude that at least one of the covariates is related to the effect size.

Goodness of fit

This covers sections [C] in Figure 20.

The statistics in this section address the question "Is able to explain *all* of the variation in effect size?" The null hypothesis is that the true effect size for each study falls directly on the regression line, and all variation about the regression line is due entirely to the fact that the observed effects (which are the ones on the plot) differ from the true effects.

The statistics presented here have the same interpretation as those presented in the prior regressions. In all cases, the statistics describe the dispersion of effects about the regression line. In Analysis-2, the regression line was horizontal – the predicted effect size for all studies was the same. In Analysis-3, the regression line was based on Dose. In Analysis-4, the regression line (actually displayed as two lines) was based on SUD. In the current analysis, the regression line is based on Dose and SUD. As such, it takes the form of two lines as in Figure 25 – there's one regression line that shows the relationship between Dose and effect size for SUD studies, and another that shows the relationship between Dose and effect size for non-SUD studies .

The standard deviation of true effects, T, is 0.0314. If we assume that most true effects will fall within two standard deviations of the regression line, then most will fall within 0.06 of the regression line. For a study that enrolls SUD patients and uses a Dose of 30 units, the predicted effect size is ____. The true effect size for that study probably falls in the range of ___ to ____.

The variance of true effects, T^2, is 0.0010. This is simply the standard deviation, squared. This is a less intuitive metric than *T*, but it has some useful statistical properties.

The *Q* statistic may provide a test of the null hypothesis that the true effect size for all studies is the predicted value. This could also be framed as "The true effect size for all studies lies precisely on the regression line."

The *Q*-value is 14.3060 with 14 degrees of freedom and a p-value of 0.4272. When we test for heterogeneity we typically use a criterion alpha of 0.10. On this basis we would not reject the null. The dispersion of effects about the regression line could be due entirely to sampling error. Another way to say this, is that the actual value of of τ (and τ 2) could be zero.

Finally, I^2 is reported as 2.14%. As always, I^2 does not tell us how much the effects vary. Rather, it provides some context for the variance of observed effects. While the plot shows some variation in observed effects about the regression lines, only about 2% of this reflects variation on true effects. If we could somehow plot the *true* effect size (rather than the observed effect size) for each study, the effects would tend to fall much closer to the regression line than they do in Figure 25.

Comparison of Model 1 with the null model

When we perform a regression analysis in a primary study we often report a statistic called R2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T^2_0 is the variance of true effects about the mean, and T^2_1 is the variance of true effects about the regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

Total between-study variance (intercept only)

When we perform a regression analysis in a primary study we often report a statistic called R2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T^2_0 is the variance of true effects about the mean, and T^2_1 is the variance of true effects about the regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

In section [D] we saw that with no covariates in the model, the variance of true effects is 0.0387. This is the same value reported in Analysis-1 and Analysis-2, and it will serve as the baseline value for computing R^2. In section [C] we saw that the variance of true effects about the regression line based on Dose and SUD was 0.0010. The difference,

$$T^2_{Explained} = T^2_{Total} - T^2_{Residual} = 0.0387 - 0.0010 = 0.0377,$$

is the variance in true effects explained by SUD. We can then estimate the proportion of variance explained by SUD using

$$R^2 = \frac{T^2_{Explained}}{T^2_{Total}} = \frac{0.0377}{0.0387} = 0.97$$

Summary

We used meta-regression to examine the relationship between Dose, SUD, and effect size.

The coefficient for Dose is 0.0092. When we control for SUD, every one-unit increase in dose is associated with an increase of 0.0093 in effect size. As Dose increases by 50 units, the expected effect size increases by 0.46. The 95% confidence interval for the coefficient extends from 0.0027 to 0.0156. The *Z*-value for a test of the null (that the true coefficient is zero) is 2.7797, and the corresponding *p*-value is 0.0054. [A]

The coefficient for SUD| Yes is −0.4492. When we control for Dose, the mean effect size in studies that enrolled SUD patients was 0.4492 lower than for studies which excluded these patients. The 95% confidence interval for the coefficient extends from −0.7321 to −0.1664, which tells us that the difference between groups probably falls in the range. The *Z*-value for a test of the null (that the true coefficient is zero) is -3.1130, and the corresponding *p*-value is 0.0019. [A]

A test of the model yields a *Q*-value of 15.5869 with 2 degrees of freedom and corresponding *p*-value of 0.0004. We conclude that the model is able to explain at least some of the variance in effect size. [B]

The variance of true effects about the regression line (T^2) is 0.0010, and the standard deviation of true effects about the regression line (T) is 0.0314. The I^2 statistic is 2.14%, which tells us that only about 2% of the observed variance about the regression line reflects variation in true effects rather than sampling error. The test for heterogeneity yields a *Q*-value of 14.3060 with 14 degrees of freedom and a corresponding *p*-value of 0.4272. The variation of observed effects about the regression line falls within the range that can be explained by sampling error alone. [C]

The R^2 analog is 0.97, which means that the model is able to explain some 97% of the variance in true effects. [E]

The relationship between these covariates and effect size is observational, not causal. The treatment-effect is related to SUD after we partial Dose, but could be due to other confounds. Similarly, the treatment-effect is related to Dose even after we partial SUD, but could be due to other confounds.

PUTTING IT ALL TOGETHER

Above, we explained the meaning of all statistics reported on the results screen. Here, we offer an example of how one might report the results of this regression. This is intended only as a starting point. Depending on what out a priori hypotheses had been, we would highlight some elements in this report and minimize others.

Summary

We used meta-regression to examine the relationship between Dose, SUD, and effect-size.

When Dose is the only covariate in the model, the coefficient for Dose is 0.0092. When we control for SUD, the coefficient for Dose is 0.0093. Thus, the relationship between Dose and effect size cannot be explained as a confound with SUD.

When SUD is the only covariate in the model, the coefficient for SUD | Y is -0.4158. When we control for Dose, the coefficient for SUD is −0.4492. Thus, the relationship between SUD and effect size cannot be explained as a confound with Dose.

The summary that follows is based on Analysis-5. As such, it reports the unique impact of each covariate, as well as the combined impact of the two covariates.

Dose

Every one-unit increase in dose is associated with an increase of 0.0093 in effect size. When we control for SUD, As Dose increases by 50 units, the expected effect size increases by 0.46. The 95% confidence interval for the coefficient extends from 0.0027 to 0.0156. The *Z*-value for a test of the null (that the true coefficient is zero) is 2.7797, and the corresponding *p*-value is 0.0054.

SUD

The coefficient for SUD| Yes is −0.4492. When we control for Dose, the mean effect size in studies that enrolled SUD patients was 0.4492 lower than for studies which excluded these patients. The 95% confidence interval for the coefficient extends from −0.7321 to −0.1664, which tells us that the difference between groups probably falls in the range. The *Z*-value for a test of the null (that the true coefficient is zero) is -3.1130, and the corresponding *p*-value is 0.0019.

The actual magnitude of the predicted effect is plotted in _____. For studies that exclude SUD patients, the effect size varies (in round numbers) from 0.25 (at the lowest dose) to 0.80 (at the highest dose). For studies that includes SUD patients, the effect size varies (in round numbers) from −0.2 (at the lowest dose) to +0.25 (at the highest dose).

A test of the model yields a *Q*-value of 15.5869 with 2 degrees of freedom and corresponding *p*-value of 0.0004. We conclude that the model is able to explain at least some of the variance in effect size. [B]

The variance of true effects about the regression line (T^2) is 0.0010, and the standard deviation of true effects about the regression line (T) is 0.0314. The I^2 statistic is 2.14%, which tells us that only about 2%

of the observed variance about the regression line reflects variation in true effects rather than sampling error. The test for heterogeneity yields a *Q*-value of 14.3060 with 14 degrees of freedom and a corresponding *p*-value of 0.4272. The variation of observed effects about the regression line falls within the range that can be explained by sampling error alone.

The R^2 analog is 0.97, which means that the model is able to explain some 97% of the variance in true effects. [E] Estimates of this statistic are not always reliable, and this value is probably an over-estimate of the proportion of variance actually explained.

The relationship between these covariates and effect size is observational, not causal. The treatment-effect is related to SUD after we partial Dose, but could be due to other confounds which were not measured, or which we could not include in the analysis since we had a limited number of studies.

Similarly, the treatment-effect is related to Dose even after we partial SUD, but could be due to other confounds. The relationship between SUD and effect size is especially susceptible to potential confounds, since the sample includes only four studies that enrolled SUD patients. In particular, we note that two of the four SUD studies used a continuous (rather than an intermittent) formulation. The fact that these studies had exceptionally low effect sizes (one was negative) could be due to the inclusion of SUD patients or to the formulation (or to both). With the data at hand, it's not possible to identify the unique impact of either variable.

Based on this data, one could argue forcefully for the development of a clinical trial where patients are randomly assigned to different doses. This would allow us to determine whether or not there is a causal relationship between Dose and effect size. We can also use the results of the meta-analysis to help in planning the clinical trial. In particular, the inclusion of SUD patients, and also the type of formulation, may have a strong impact on the effect size. Therefore, the study would be better able to assess the impact of Dose if we exclude SUD patients and employ only an intermittent formulation.

META-REGRESSION IS OBSERVATIONAL

Our goal in this chapter is to explain that the relationships identified by meta-regression are observational rather than causal. We'll start with a primary study to provide context, and then show how the same ideas can be extended to a meta-analysis.

Regression in primary study

1. Consider a study where participants are assigned at random to either treatment or placebo. If the treated group performs better than the placebo group, we would conclude that the treatment is responsible for the difference. We can do this because the randomization (if it worked properly) ensures that the subjects in the two groups are similar in all respects except for the treatment.

2. By contrast, consider a study where we don't assign people to a treatment but rather classify people based on existing characteristics. For example, we locate people who are taking (or not taking) a specific drug, and find that those in the treated group perform better. This *could* be because of the drug, but could also be because people who chose to take the drug also tend to exercise more than those who did not take the drug, and it's the exercise (rather than the drug) that's responsible (in whole or in part) for the better outcome. We can report that patients taking the drug had a better outcome than the others, but we cannot say that the drug is responsible for the difference.

3. That said, we might use multiple regression to try and isolate the impact of the drug. That is, we would see if a prediction model that includes drug in addition to exercise, is a better predictor of outcome than a model which includes the drug alone. This approach requires that two conditions be met. First, we need a sufficient number of subjects. As a rule-of-thumb, researchers often require at least ten subjects per covariate. Second, the covariates cannot be highly correlated with each other. In this example, if all those taking the drug also exercised, it would not be possible to isolate the impact of drug. Finally, even if all of these conditions are met, and it turns out that those taking the drug performed better even when exercise is held constant, we still could not *conclude* that the drug is responsible for the better outcome, since we may have overlooked other confounds.

Regression in a meta-Analysis

1. Suppose that all studies in the meta-analysis had been coordinated in advance by a consortium of researchers who enrolled twenty hospitals and then randomly assigned each hospital to test either high-dose vs. placebo or low-dose vs. placebo. In the meta-analysis we compute the mean effect size (treated vs. control) separately for the low-dose studies and for high-dose studies. If the effect size is larger for the former, we would conclude that the dose is responsible for the difference. We can do this because the randomization of hospitals to dose (if it worked properly) ensures that the studies in the two sub-groups are similar in all respects except for the dose.

2. By contrast, consider the case where we locate twenty studies in the literature and classify each as low-dose vs. placebo or high-dose vs. placebo. If it turns out that the effect size is larger in the former, this *could* be because a high-dose is more effective than a low-dose, but it could also be because the high-dose studies differed from the low-dose studies in other ways. For example, it's possible that researchers working with severe cases of ADHD tended to employ a high dose, while those working with more moderate cases tended to employ a low dose. And, it could be the severity of the condition, rather than the dose, which is responsible for the difference – severe cases may have been more likely to improve. Therefore, we can report that the effect size was larger in the subgroup of studies that employed a high dose, but we cannot say that the dose is responsible for the difference. Put simply, we are calling these the "High Dose" studies but maybe a more appropriate label would be the "Severe cases" studies.

3. That said, we might use meta-regression to try and isolate the impact of the dose. That is, we would see if a prediction model that includes dose in addition to severity is a better predictor of effect size than a model which includes severity alone. This approach requires that two conditions be met. First, we need a sufficient number of studies. As a rule-of-thumb we might require at least ten studies per covariate. Second, the covariates cannot be highly correlated with each other. If all the studies that employed a high-dose also enrolled severe cases, it would not be possible to isolate the impact of dose. Even if these conditions are met, and it turns out that the effect was larger in the high-dose studies when severity is held constant, we still could not *conclude* that the relationship is causal, since we may have overlooked other confounds.

In meta-regression, virtually all relationships are observational

For both primary studies and for meta-analysis we outlined two basic cases – the randomized case and the observational case. In primary studies both cases exist in practice, but in meta-analysis the randomized case that we outlined as case (1) would only be found in a multi-center trial that is being analyzed as a meta-analysis. Outside of that case, virtually all subgroup comparisons or regressions that we perform in a meta-analysis will be observational.

Impact of risk factors is observational even if the studies are RCTs

To be clear, the relationship between any risk factor and effect size is observational *even if each study in the meta-analysis is a randomized controlled trial (RCT)*.

In the ADHD example, every one of the studies is an RCT. Nevertheless, the randomization protects against confounds only when we are comparing drug vs. placebo, because this is where the randomization takes place. It does not protect against confounds when we compare Low-Dose studies vs. High-Dose studies because this distinction is based on non-randomized differences. Therefore,

- If the meta-analysis shows that the main effect (drug vs. placebo) is statistically significant, we *can conclude* that the drug is probably responsible for this difference.

- If the subgroup analysis (or regression) shows that the effect is stronger for high-dose studies than for low-dose studies, we *cannot conclude* that the dose is probably responsible for this difference.

When is the regression likely to be informative?

Earlier, we noted that a regression will only be informative if we have enough data to isolate the unique impact of the various factors. We use three cases to outline how this might work. For purposes of this example we'll assume that the only factors we expect to impact the effect size are formulation (the treatment was delivered either continuously or intermittently), SUD (the study included or excluded substance abusers) and Dose.

Case A | The ideal case

In the ideal case, three things are true. (1) We have a large number of studies in each subgroup, (2) we have measures for all the putative risk factors, and (3) the risk factors are not highly confounded with each other.

In the present example this would mean that (1) we have at least 15 studies (the number is somewhat arbitrary) that employed a continuous formulation and 15 that employed an intermittent formulation, (2) we believe that the only factors which affect the impact of methylphenidate are formulation, substance abuse, and dose, and (3) the two subgroups (consistent and intermittent formulation) each include a reasonable proportion of studies with low dose/high dose and with/without SUD.

In this ideal case we would be able to obtain a reasonably precise estimate of the mean effect for the two regimens (say, 0.10 for continuous vs. 0.70 for intermittent). We could then partial the impact of dose and substance abuse. If the difference remained, we would have reason to believe that formulation does have an impact on the effect size.

It's still possible that there is an unknown risk factor which is confounded with formulation, and so we cannot conclude with certainty that there is a causal relationship. However, the observational study tells us that there *probably* is a causal relationship, and we might then plan a randomized trial to test this hypothesis.

By contrast, consider the two following cases.

Case-B | Confound with other risk factors

Consider the same situation as Case-A, where we have fifteen studies on each formulation, but this time assume that formulation is highly confounded with another risk factor. That is, (almost) all of the studies which employed a continuous formulation also employed a low dose. Or, (almost) all of the studies which employed a continuous formulation also enrolled patients with SUD. If the continuous-formulation studies prove to have a lower effect size, there's no way to determine if this is due to the formulation itself or due to the other risk factor. The partialing process doesn't work in this case.

Case-C | Small number of studies

When we have only a small number of studies the issue outlined as Case-B will invariably be a problem. Concretely, if we have only a few studies that employed continuous formulation, then it follows that we do not have a substantial number of studies in this group that included SUD patients and a substantial number that excluded them.

In Case-B we are primarily concerned with factors that are systematically related to the risk factor of interest. For example, there might be a reason that researchers who employ a continuous formulation also tend to enroll patients with SUD. We are less concerned with factors that are not systematically related to the risk factor of interest, since these factors will usually be distributed across both types of formulation.

By contrast, when we have only a small number of studies in either subgroup (Case-C) we cannot rely on the play of chance to help protect against other risk factors. For example, suppose that the treatment is less effective in older patients. If we have 15 studies in each subgroup and there is no systematic relationship between subgroups and age, then it's likely that the mean age will be comparable across subgroups. However, with only two studies in a subgroup, it's possible that the mean age in these two studies will be unusually low or high simply by the play of chance. In that case, if age is indeed related to effect size, this could have an impact on the mean effect size in each subgroup.

This is a general problem

To be clear, this problem is not unique to regression. We face *precisely* the same problem when we compare the effect size in two or more subgroups of studies. There, the problem is often ignored, but we need to be aware of it.

Consider an analysis that compares (for example) three subgroups. One of these subgroups includes five studies, another includes ten studies, and a third includes one study. Suppose this third subgroup is labeled "Older patients" and demonstrates an exceptionally large effect size. The assumption (implicit and/or explicit) is that the treatment is most effective in older patients. But if this subgroup includes only one study, it's likely that this study differs from the others in ways other than the patients' age. The dose may have been higher; the intervention may have been more intensive; the study may have been performed at a hospital that offered unique kinds of help. Therefore, whenever a subgroup includes a small number of studies, there is a real concern that the magnitude of the effect size in this subgroup could be due to a confound.

It's relatively simple to elucidate this problem (as we did in the previous paragraph) when we are working with subgroups. The problem may be harder to explain when we move to regression, yet the problem is *precisely* the same. In a later chapter we show how we can use regression to compare subgroups, as we get *precisely* the same results if we use a subgroups comparison or a regression analysis. Behind the scenes, both are doing exactly the same thing.

In the current example, one of the covariates we looked at is SUD, and we found that the inclusion of SUD patients is associated with a lower effect size. If we had a large number of SUD studies – say, 15 with SUD and 15 without SUD, we would still be concerned that the lower effect for the SUD studies was due to a systematic confound. For example studies that enrolled SUD patients might tend to have poor compliance. However, if we know the field, we would probably be able to include the potential confounds in the analysis and see if the SUD effect remained even after we controlled for these confounds.

By contrast, when we have only four studies that enrolled SUD patients, we need to be concerned not only about systematic confounds, but also about random confounds. For example, it's possible that some of these studies happened to enroll patients who were unique in some other way that affected

the impact of the drug. Since this is a random confound, we wouldn't think to control for it, and may not have measured it.

There is no way around this basic issue, except to be aware of it and respect it. Throughout this volume we say, for example, that a higher dose of the drug is associated with a higher effect size – not that it *causes* a higher effect size, but that it's *associated* with a higher effect size.

As noted, this issue is not unique to meta-regression. We may face precisely the same problems when using multiple regression in a primary study. However, in a primary study we often have enough cases so that (a) we can do a good job of controlling for systematic confounds and (b) don't need to be too concerned about random confounds. In meta-regression, when we have relatively small numbers of studies, we need to be concerned about both of these issues.

When the regression suggests that a variable is related to the effect size, and that this relationship remains when we partial all known confounds, the next step might be to plan a new primary study that can actually test the relationship. In the current example, this might require that patients are randomized to different doses of the drug.

SHOW WHAT HAPPENS TO CURVILINEAR WHEN WE ADD SUD. TOO FEW CASES TO DO THIS PROPERLY, BUT AS AN EXAMPLE OF WHAT WE DON'T KNOW BECAUSE OF TOO FEW CASES

Of course, it's possible that the drug really is more effective at low and high doses. But it's possible (and probably more likely) that this is due to a confound. The issue here is that there are only a handful of studies at the middle dose, and if a few of these happen to have low effects due to some random confound, we might not know it. If we partial _____.

Meta-regression is observational

- If all studies in the analysis are randomized controlled trials (RCTs), then the protection afforded by the randomization process in the primary studies carries over to the main effect in the meta-analysis. If the effect in each study is due to the intervention, we can say that the summary effect in the meta-analysis is probably due to the intervention.

- However, even if the individual studies are randomized trials, once we move beyond the goal of reporting a summary effect and proceed to perform a subgroup analyses or meta-regression, we have moved out of the domain of randomized experiments, and into the domain of observational studies. If the studies that employed a higher dose showed a higher effect than those which employed a lower dose, this *could be* because a higher dose is more effective. But it *could also be* because the studies which employed a higher dose worked with populations where the treatment worked better for a variety of reasons.

- That said, in primary observational studies, researchers sometimes use regression analysis to try and remove the impact of potential confounders. This is not a perfect solution since there may be other confounders of which we are not aware, but this approach can help to isolate the impact of specific factors and generate hypotheses to be tested in randomized trials. The same holds true for meta-regression.

- This approach is potentially useful only when there are enough studies to isolate the unique impact of each factor. Many meta-regressions are based on relatively small numbers of studies, and so it may not be possible to adjust for potential confounds.

- Even when there are enough studies to adjust for known confounds, we cannot be certain that we've identified all possible confounds. Therefore, we can't use this approach to prove a causal relationship. Rather, we can use it to identify factors which are probably related to the effect size, and then test these in new primary studies.

TESTING THE NULL

In some fields of research, there is a longstanding controversy over the use of null-hypothesis significance testing (NHST) on the one hand, versus effect-size estimation, on the other. Hundreds of papers and several books address this controversy in the context of primary studies. Our goal here is to show how the same issues play out in the context of a meta-analysis.

- Under the NHST approach we pose the null hypothesis that some parameter is zero. Then we compute a test statistic and a corresponding p-value. A p-value of 0.05 (for example) tells us that if the null hypothesis is true, we would see an effect size as large as we did (or larger) only 5% of the time. If the observed p-value is closer to zero than the criterion (alpha), we reject the null hypothesis and conclude that the actual parameter is probably not zero.

- Under the effect-size estimation approach we report the size of the effect along with a confidence interval that speaks to the precision of the estimate. For example, we might report that the treatment increases the mean score by 50 points with a confidence interval of 35 to 65 points.

The decision to use one approach or the other is typically determined by tradition, but in fact should be determined by the goals of the analysis. NHST is appropriate for analyses where we actually have an interest in testing the null hypothesis. Effect size estimation is appropriate for analyses where we want to assess the magnitude of the effect.

An example where NHST would be appropriate

Consider a study that compares the impact of homeopathic medicines vs. placebo. A homeopathic medicine is one in which a compound is placed in a solution of water, and then repeatedly diluted until none of the original compound actually remains in the solution. To state this clearly, the bottle of homeopathic medicine cannot be chemically distinguished from water – it contains zero molecules of the original compound. The idea is that the water "remembers" what had been there before, and therefore retains some medicinal properties (see box).

While logic dictates that the homoeopathic remedy cannot be more effective than water, homeopathy is nevertheless a multi-billion dollar industry. Therefore, the question of whether or not the remedy has any effect is of interest. Studies have shown that people who are treated with homeopathic medicines have a better outcome than those not treated, but there are two possible reasons for these findings. (A) The better outcome could be due to entirely to the patient's interaction with the homeopath, or (B) the homeopathic solutions could actually have some healing properties.

To test these competing hypotheses we devise a double-blind randomized study. In this study, all patients will be treated by homeopaths. In half the cases, the homeopath will provide the actual homeopathic remedy. In the other half, the homeopath will provide water. Critically, the assignment is random and the homeopath is not aware of which solution he or she is providing. The null hypothesis is that the two treatments are equally effective – that the proportion of patients who report an improvement in symptoms, or the mean outcome after a period of time, will be identical in the two groups.

People who believe that the entire benefit associated with homeopathic medicine is due to the interaction of the patient with the homeopath and/or the patient's expectations, expect that the null hypothesis is true. Those who believe that the water has memory, expect that the null hypothesis is false. These are two radically different visions of homeopathy, and a valid test of the null hypothesis will provide strong evidence one way or the other. In this example, *it really is the null hypothesis that we care about*, and that we should focus on, in the analysis. If we can reject the null hypothesis – if we can prove that the homeopathic solution is better than the water by *any* amount, the finding would radically alter our view of the world.

An example where effect-size estimation is appropriate

In the vast majority of analyses in medicine and social science, we don't really care about the null hypothesis. Rather, we care about estimating the magnitude of the effect.

The ADHD analysis serves as a case in point. Consider any single study where patients are randomized to receive either methylphenidate or placebo. The reason we are studying the impact of methylphenidate is to determine whether or not it's a clinically useful intervention. The answer depends on the magnitude of the effect, and not on whether or not the impact is precisely zero. The only information provided by a test of the null (and a significant *p*-value) is that the effect is not precisely zero. By contrast, the information provided by an estimate of the effect size is that the treatment increases scores by a given amount – in our example, by one-half a standard deviation. This is the information we need in deciding whether or not the treatment is clinically useful.

There is also a second reason to avoid NHST in these kinds of analyses. The null hypothesis asserts that the difference in means for treated vs. control is 0.0000000000, and we can safely assume that this null hypothesis is false. Even if the impact of methylphenidate is too small to be of any clinical import, as a chemical it will almost certainly have *some* effect on scores, even if that effect is 0.0000000001. If we reject the null then we are simply confirming the obvious. If we fail to reject the null, it's because the sample size is too small. In that case, the null hypothesis is known to be false, and so the test is pointless.

The homeopathy example is different, because (for those who believe that the null hypothesis is true) both conditions are exactly identical to each other – both groups are getting pure water. In this case, we don't want to know if the effect size is trivial, modest, or large. Rather, we want to know whether or not it is precisely 0.000000000.

The first two reasons for avoiding NHST assume that the results are being interpreted properly. A third reason for avoiding NHST is that when researchers focus on a test of significance, the results are often misinterpreted. A significant p-value is interpreted to mean that the treatment is clinically useful, and a non-significant p-value is interpreted to mean that the treatment is not clinically useful. In fact, the p-value cannot be interpreted in this way.

The p-value is a function of two elements – the magnitude of the effect and the size of the sample. When the p-value is significant, it's possible that the effect size is clinically important, but it's also possible that the effect size is trivial, and the sample size is large. Conversely, when the p-value is not significant, it's possible that the effect size is trivial, but it's also possible that the effect size is clinically important, but the sample size is small.

If our goal really is to test the null hypothesis (as it was in the homeopathy example), then using the p-value makes sense. If we can reject the p-value, then we have evidence that the effect size is not precisely zero. By contrast, if our goal is to determine whether or not a treatment is clinically useful, then it's critical to estimate the size of the effect, and (as a separate matter) the precision of the estimate. That's what we do when we focus on the effect size.

For these reasons, in the vast majority of primary studies in medicine and social science, the effect-size approach is a better match for our goals than NHST. The same holds true when we move on to meta-analyses in these fields. We will use the ADHD meta-analysis as a case in point.

The main effect

One goal of the analysis was to assess the impact of the drug on cognitive function. We have the option of using NHST, or of focusing on the magnitude of the effect size. In the simple analyses these are addressed by the Z-value (and p-value) for the former, and by the mean effect size with its confidence interval, for the latter.

If we use the NHST approach we pose the null hypothesis, that the drug has no effect on cognitive function. A test of this null hypothesis yields a Z-value of _____ with a corresponding p-value of _____. So we reject the null, and conclude that, in the universe of studies from which these studies were sampled, the impact of the drug is not zero. But this does not tell us what the impact is. The substantive impact could be trivial or it could be substantial.

One could also argue that the null hypothesis is meaningless in this context. It seems very unlikely that the impact of these drugs on cognitive function is *precisely* 0.0000000, and *that* is what the null hypothesis is testing.

If we use the effect-size approach, we report that the mean effect is 50 points – on average, the drug increased cognitive function by one-half a standard deviation. The actual value (in the universe of studies from which these studies were sampled) is at least 35 points and possibly as much as 65 points. So, we know the size of the effect, and also the precision of the estimate.

In general, we don't care whether or not the mean effect size is precisely zero. Rather, we want to know the magnitude of the effect size. Therefore, we should focus on the magnitude of the latter.

The impact of covariates

One goal of the analysis was to assess the relationship between Dose and effect-size. We have the option of using NHST, or of focusing on the magnitude of the effect size. In the regression these are addressed by the Z-value (and p-value) for each covariate (for the former) or the coefficient and its confidence interval (for the latter).

If we use the NHST approach we pose the null hypothesis, that the effect size has no (linear) relationship to Dose. A test of this null hypothesis yields a z-value of ____ with a corresponding p-value of ____. So we reject the null hypothesis and conclude that, in the universe of studies from which these studies were sampled, the relationship is not zero. Higher Doses are associated with higher effects. But this does not tell us anything about the size of the relationship.

One could also argue that the null hypothesis is meaningless in this context. It seems very unlikely that the impact of the relationship between Dose and effect size is *precisely* 0.0000000, and *that* is what the null hypothesis is testing.

If we use the effect-size approach, we report that a 50-unit increase in Dose corresponds to a 45 point increase in effect size. In the universe of studies from which these studies were sampled, the actual increase is probably at least ___ points and possibly as much as _____ points. So, we know the size of the effect, and also the precision of the estimate.

In general, we don't care whether or not the relationship between dose and effect size is precisely zero. Rather, we want to know the magnitude of the relationship. Therefore, we should focus on the magnitude of the coefficient.

Heterogeneity

The two prior examples (for the main effect and the covariates) are simple extensions of the issues we typically see in primary studies. In meta-analysis we also need to address the assessment of heterogeneity. We'll show how we address that for a simple analysis and also for a regression.

Heterogeneity in a simple analysis

One goal of the analysis was to assess the dispersion of effects about the mean. We have the option of using NHST, or of focusing on the magnitude of dispersion. These are addressed by the *Q*-value, *df*, and and *p*-value (for the former) or the variance, standard deviation and prediction interval (for the latter).

Using NHST we pose the null hypothesis that the true effect size (the effect size that we would see if we could remove all sampling error) is precisely the same for all populations in the universe from which these studies were sampled. A test of this null hypothesis yields a Q-value of _____ with 16 degrees of freedom, and a corresponding p-value of _____. We reject the null hypothesis that the true effect size is the same in all studies. However, this tells us nothing about how much the effects vary.

One could also argue that the null hypothesis is meaningless in this context. The null hypothesis asserts that the impact of the drug will be exactly the same for all populations in the universe from which we are sampling. To suggest that the impact will be exactly the same in two populations (let alone all the relevant populations) is simply not plausible. Additionally, since the details of the intervention varied across studies, the null hypothesis could only be true if these variations (the precise dose, duration, and so on) had zero impact on the effect size. Again, this is not plausible.

If we use the effect-size approach we focus on the extent of heterogeneity. In this case we would report that the standard deviation of true effects is around 0.20, and that the true effect size in most populations falls in the range of 0.10 to 0.90. This addresses the issue that we care about, which is "How much do the effects vary?"

In general, we don't care whether or not the effect size varies at all across studies. Rather, we want to know how much it varies. This is addressed by T and the prediction interval.

Heterogeneity in a regression

One goal of the analysis was to assess the dispersion of effects about the regression line. We have the option of using NHST, or of focusing on the magnitude of dispersion. These are addressed by the *Q*-value, *df*, and and *p*-value (for the former) or the variance, standard deviation and prediction interval (for the latter).

Using NHST we pose the null hypothesis that, in the universe from which these studies were sampled, the true effect size is precisely the same for all populations that share the same Dose and SUD, A test of this null hypothesis yields a Q-value of _____ with 16 degrees of freedom, and a corresponding p-value of _____. We cannot reject the null hypothesis that the true effect size is the same in all studies. With the same Dose and SUD.

One could also argue that the null hypothesis is meaningless in this context. The null hypothesis asserts that the impact of the drug will be exactly the same for all populations in the universe with the same Dose. To suggest that the impact will be exactly the same in two populations with the same Dose (let alone all the relevant populations) is simply not plausible. Additionally, since the details of the intervention varied across studies, the null hypothesis could only be true if these variations (the duration, and so on) had zero impact on the effect size. Again, this is not plausible.

If we use the effect-size approach we focus on the extent of heterogeneity. In this case we would report that the standard deviation of true effects is around 0.15, and that the true effect size in most populations falls within 0.30 of the predicted effect size. If the Dose is 30 units, the true effect size will usually fall in the range of ____ to ____. If the Dose is 80 units, the true effect size will usually fall in the range of ____ to _____. This addresses the issue that we care about, which is "How much do the effects vary?"

In general, we don't care whether or not the effect size varies about the regression line *at all*. Rather, we want to know how much it varies. This is addressed by T and the prediction interval.

In context

At the outset of this chapter we referred to a "controversy" about the relative merits of NHST vs. effect-size estimation. Some would say that the controversy has been resolved (at least in such fields as medicine, education, and social science), with a widespread acknowledgement that we should focus on effect-size estimation. While we obviously concur, we recognize that NHST continues to play an important role. In particular, some governmental agencies will base their decision to approve a treatment on NHST rather than effect-size estimation. As a practical matter, then, NHST will continue to play a role for the foreseeable future.

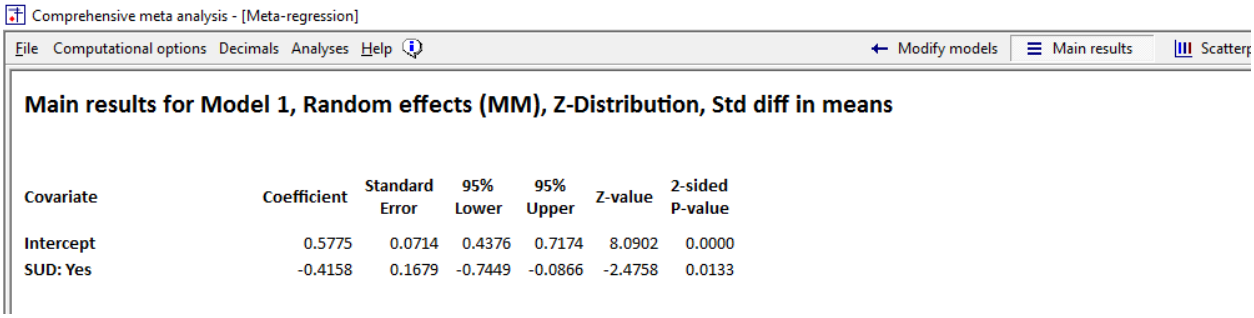
Fortunately, there's no real harm in presenting a test of the null hypothesis as long as these results are interpreted correctly. As noted above, when a p-value is presented in isolation, people tend to push them into use as a surrogate for the effect size. By contrast, if one presents a test of the null hypothesis *and also* an estimate of the effect size, there is less concern that the p-value will be misused in this way.

Note that this chapter addresses the idea that we should focus on effect sizes rather than NHST when discussing the *results* of a meta-analysis. When we consider what kind of data to use as *input* to the analysis, it's clear that we must use effect sizes rather than p-values.

Explain why test for heterogeneity is based on fixed-effect weights (actually random-effect with T=0)

Categorical variables

When you plot by a categorical variable the program uses a different format to display the plot. For example, SUD is a categorical variable that is coded "No" for studies that excluded patients with substance-abuse problems and "Yes" for studies that included these patients. Figure 26 shows the results of the regression and Figure 27 shows the plot.



Main results for Model 1, Random effects (MM), Z-Distribution, Std diff in means

Covariate	Coefficient	Standard Error	95% Lower	95% Upper	Z-value	2-sided P-value
Intercept	0.5775	0.0714	0.4376	0.7174	8.0902	0.0000
SUD: Yes	-0.4158	0.1679	-0.7449	-0.0866	-2.4758	0.0133

Figure 26 | Regression | Main results | Random-effects

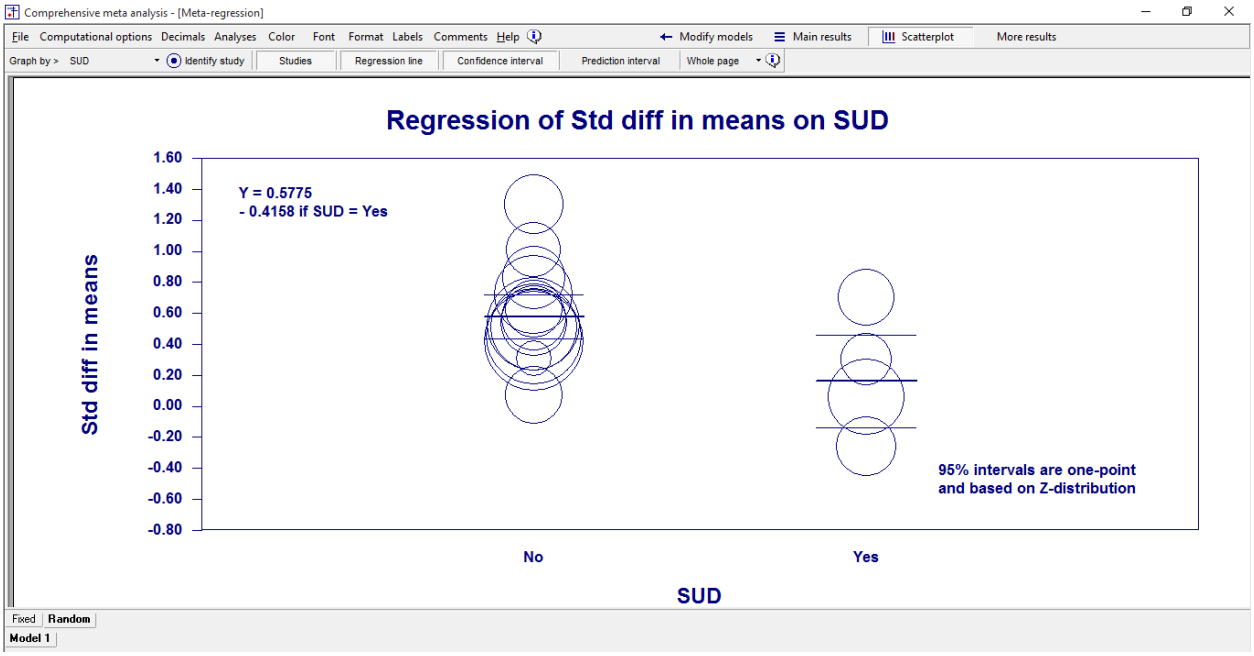
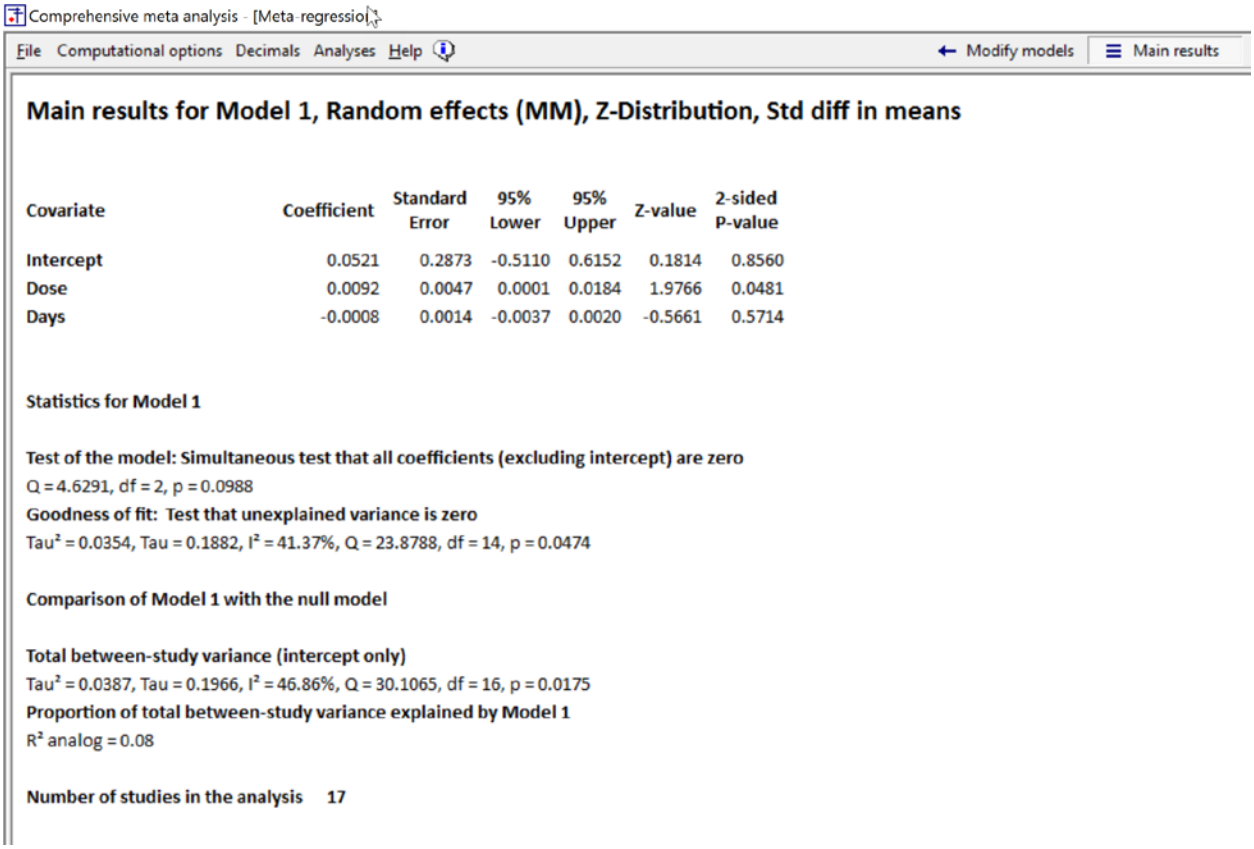


Figure 27 | Regression | Plot | Categorical covariate

The plot shows one column for every category (No, Yes). Rather than link the column means with a regression line, the program plots the mean and confidence interval for each column separately. All other options work the same way as they do for continuous covariates.

Plots when there are two or more covariates



If there is more than one covariate in the prediction equation, select the covariate to be plotted. Either click “Graph by on the toolbar” [A] or Right-click on the covariate name in the plot [B]. In this example the covariates are Dose and Duration. In Figure 28 we’ve plotted by Dose, and in Figure 29 we’ve selected Duration.

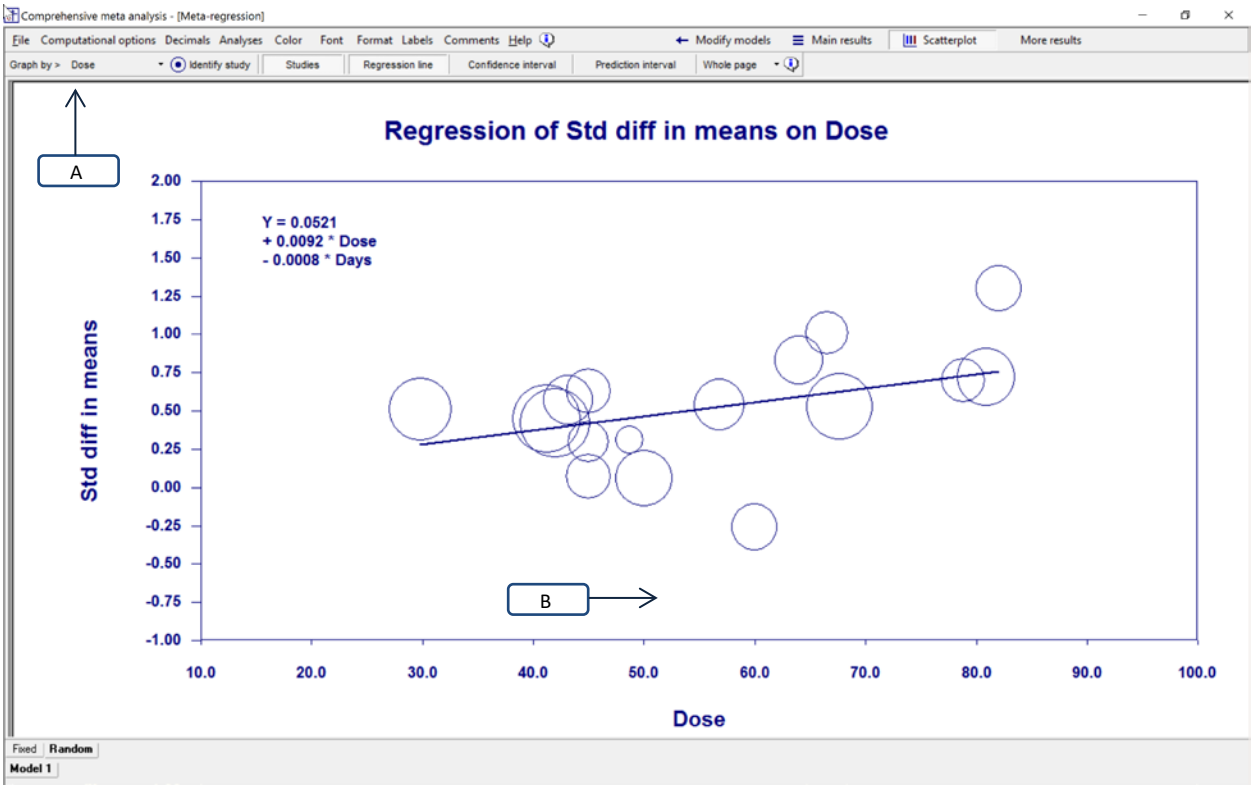

Comprehensive meta analysis - [Meta-regression]
File Computational options Decimals Analyses Color Font Format Labels Comments Help
Modify models
Main results
Scatterplot
More results
Graph by > Dose
Identify study
Studies
Regression line
Confidence interval
Prediction interval
Whole page
Regression of Std diff in means on Dose
A
Y = 0.0521
+ 0.0092 * Dose
- 0.0008 * Days
Std diff in means
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
B
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0
Dose
Fixed
Random
Model 1

Figure 28

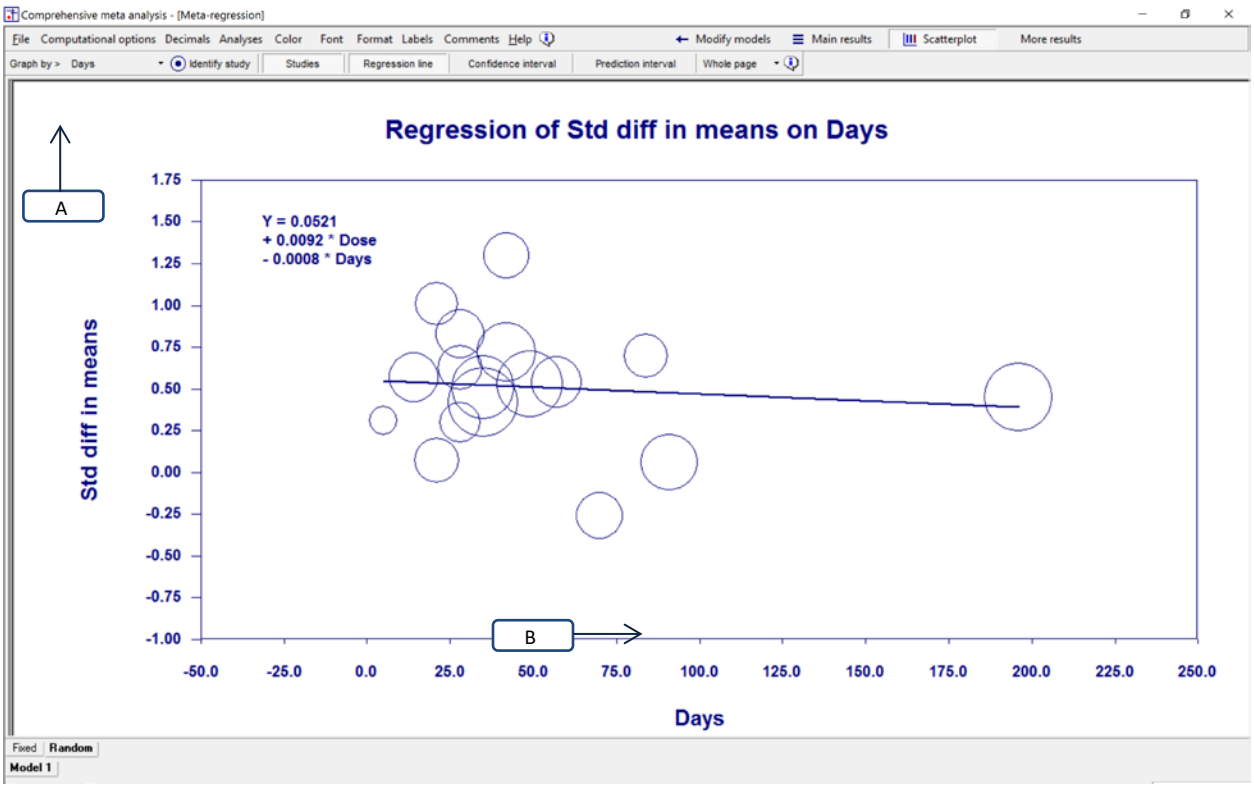


Figure 29

The program will plot effect size as a function of the selected covariate, at the mean of the covariates which are being partialled. This is a simple mean, where all studies are given equal weight.

In Figure 28 we've plotted effect size as a function of Dose. This plot is based on the mean Duration, which is _____ days. Similarly, in Figure 29 we've plotted effect size as a function of Duration. This plot is based on the mean Dose, which is 55.68 units.

The same idea holds true if the prediction equation includes a categorical covariate. For example, consider a regression where the covariates are Dose and SUD. Recall that SUD is a categorical variable where studies are classified based on whether or not they enrolled patients with a substance-abuse problem. The dummy variable corresponding to SUD is coded 0 (No) or 1 (Yes).

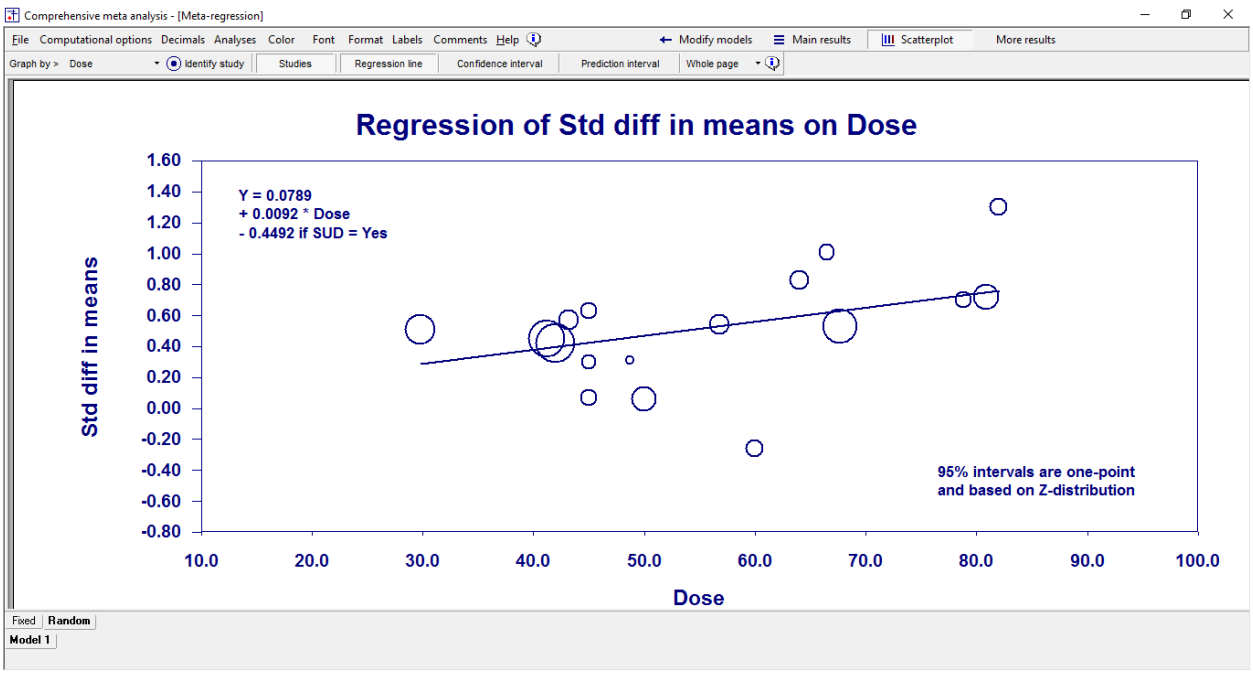


Figure 30

Figure 30 show a plot of effect size by Dose with SUD partialled. To construct the plot, we assign a value of 0.2350 to SUD, which is the proportion of studies (4/17) that enrolled SUD patients. As such, the regression line corresponds to a hypothetical study that has a 23.5% likelihood of enrolling SUD patients.

Similarly, Figure 31 show effect size as a function of SUD, with Dose partialled. Here, the plots for both SUD (No) and SUD (Yes) are based on the mean Dose.

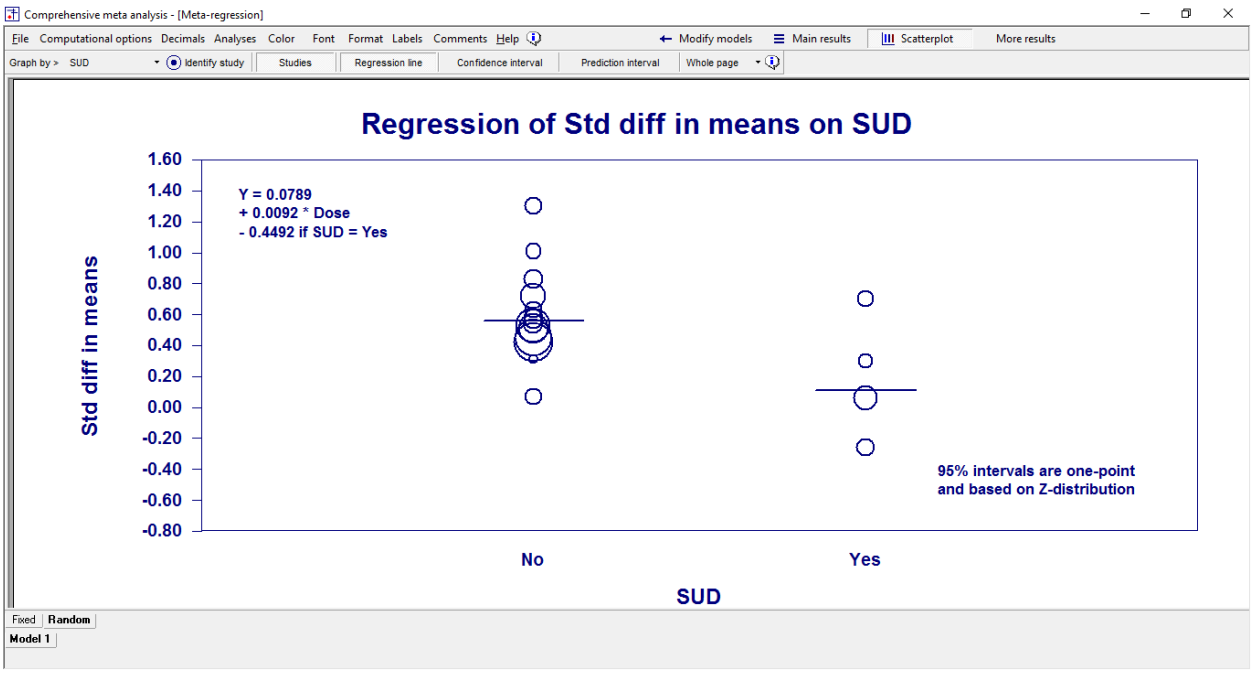


Figure 31

DIAGNOSTICS

Move, along with covariance and correlation

Regression diagnostics are designed to be simple checks that reveal important features of the data and the regression model fitted to that data. However, the multivariate situation is complex, and diagnostics are typically imperfect.

For example, consider the important feature of colinearity (correlations) among predictors. It is well known that colinearity can degrade the quality of estimates of regression coefficients by increasing their sampling uncertainty. This can occur when two predictors are highly correlated but independent of the others, or when there is a high multiple correlation among predictors (when one predictor is almost a linear combination of several others). These two situations have different implications for the quality of regression estimates. In the former case, only the two coefficients corresponding to the correlated predictors may be poorly estimated. In the latter case, the impact of colinearity may affect more coefficients. A diagnostic designed to reveal colinearity, in general, may be not be able to distinguish between the two types of colinearity. On the other hand, producing diagnostics tailored to a myriad of possible special cases increases the complexity of the suite of diagnostics, defeating the purpose of simple checks on data and the regression model.

We have implemented a set of diagnostics that have proven most useful in regression problems generally and adapted them to meta-regression. All of these diagnostics are related, in that they are different ways of looking at the extent to which the data associated with a study is inconsistent with the meta-regression model that fits the other studies. They approach the problem in different ways however.

- The leverage and Cook's distance focus on the impact of a study on the estimated regression coefficients.
- The residual and studentized residual focus on the difference between fitted (predicted) effect sizes based on all the data and the observed effect size in each study.
- *DF*ITTS and the jackknifed residual focus on the difference between fitted values of each effect size when the regression coefficients are estimated with and without a particular study.

Because these diagnostics are closely related to each other, it is not surprising that the same studies may be flagged by several of the diagnostics as having high impact or influence. In fact, it more surprising (but not impossible) when a study is flagged by only one of them.

The diagnostics should not be used by themselves to exclude studies from inclusion in a meta-analysis. The diagnostics are intended to help identify studies that have substantial impact on the estimated regression coefficients. The reference values we have given are not intended to be used like critical values in a significance test, but as criteria for further evaluation. Just because a study has high impact on the analysis does not make it incorrect. However, it is useful to know that a certain study has (or a few studies have) substantial impact on the results. In such cases it is crucial to be sure of the integrity of the studies with high impact.

It is also important to know that the impact of a study may change when the set of covariates in the meta-regression is changed or the set of studies is changed (e.g., when a subset of studies is examined).

A study that has high impact may have much less impact when a certain covariate is removed from the covariate set or a certain study is removed from the dataset.

To navigate to the diagnostics screen

- Run regression
- Click More > Diagnostics
- Select Fixed or Random at the bottom

The specific columns in the diagnostics screen are as follows

Observed value

This is simply the observed effect size

The Predicted Value

The predicted (fitted) value, $\hat{T}_i$, for the i^{th} study is the value obtained from using the estimated regression coefficients b_0, b_1, …, b_p and the covariate values for the i^{th} study x_{i1}, …, x_{ip} to compute the value of the effect size predicted for that study by the regression model

$$\hat{T}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} \ . \tag{1.15}$$

The Residual

The (unstandardized) residual value, e_i, for the i^{th} study is the difference between the observed value and the fitted value

$$e_i = T_i - \hat{T}_i \ . \tag{1.16}$$

If $e_i = 0$, the fitted value and the observed value are identical (the fitted value is exactly on the regression line or plane), but if e_i is far from 0, the predicted value is not close to the observed value.

In meta-analysis, effect sizes and their fitted values from different studies can have very different sampling uncertainties (standard errors). This makes it difficult to interpret differences in the magnitude of residuals from different studies. Standardized or jackknifed residuals attempt to address this problem of comparability by dividing the residual by its standard error.

Studentized Residual

The studentized residual value, es_i, for the i^{th} study is the residual divided by its standard error

$$es_i = \frac{e_i}{SE\left(e_i\right)} \ . \tag{1.17}$$

The standard error of es_i is given by

$$SE\left(e_i\right) = \sqrt{\frac{1-h_i}{w_i}} \ , \tag{1.18}$$

where w_i is the weight given to the i^{th} effect size in the analysis, s^2 is the weighted residual mean square, and h_i is the leverage of the i^{th} effect size. Therefore the residual divided by its standard error (the i^{th} studentized residual) is

$$ess_i = e_i \sqrt{\frac{w_i}{1-h_i}} \tag{1.19}$$

It is important to note that the standard error of the residual depends on both the residual variance which is determined by the conditional variance of the estimate (and the random effects variance component in random effects models) and the configuration of predictors (including the values for the i^{th} study). Studentized residual es_i is on a standard scale, so the values from different studies are more comparable than those of the unstandardized residuals (the e_i). If the regression model is correctly specified, the es_i have approximately a normal distribution with unit standard deviation, so that es_i values greater than 2 in absolute value occur only about 5% of the time by chance and values greater than 2.5 are quite unusual. The actual sampling distribution of es_i will often be closer to Student's t-distribution with $k - Q$ degrees for freedom, so slightly larger reference values (than 2 and 2.5) may be appropriate for judging extremeness of residuals when $k - Q$ is small.

Jackknifed Residual

The jackknifed residual ej_i, is similar to the studentized residual in that it is standardized. However the jackknifed residual is the difference between the observed effect size in the i^{th} study and the fitted value of the i^{th} study computed with the i^{th} study deleted from the dataset. That is,

$$ej_i = \frac{T_i - \hat{T}_{(i)i}}{SE\left(T_i - \hat{T}_{(i)i}\right)} \ , \tag{1.20}$$

where $\hat{T}_{(i)i}$ is the fitted value of the i^{th} study computed from all other studies except the i^{th} study. To be precise

$$\hat{T}_{(i)i} = b_{(i)0} + b_{(i)1}x_{i1} + \cdots + b_{(i)p}x_{ip} \ , \tag{1.21}$$

where $b_{(i)0}$, …, $b_{(i)p}$ are the regression coefficients estimated with the i^{th} study removed from the dataset.

The jackknifed residual is designed to better reveal cases where the i^{th} study does not fit the same model as the other studies. By removing the (potentially distorting impact of the i^{th} study from the computation of the regression coefficients, the jackknifed residual sometimes makes it easier to see how different an observed effect size is from what is expected if that study fit the meta-regression model that is appropriate for all of the other studies.

Let the weight of the i^{th} study computed using the variance component estimate with i^{th} study removed be denoted $w_{i(i)}$, then the i^{th} jackknifed residual is equivalent to

$$ej_i = e_i \sqrt{\frac{w_{i(i)}}{1-h_i}} \ . \qquad (1.22)$$

The sampling distribution of the jackknifed residual is similar to that of the studentized residual (approximately normal) and similar reference values for judging extremeness are appropriate. The actual sampling distribution of ej_i will often be closer to Student's *t*-distribution with $k - Q - 1$ degrees for freedom, so slightly larger reference values (than 2 and 2.5) may be appropriate for judging extremeness of residuals when $k - Q - 1$ is small.

Leverage

Leverage is a diagnostic that reveals how much potential influence a particular study can have on the result of the meta-regression. Let h_i be the leverage of the i^{th} study. The values of the leverage are always between zero and one inclusive, that is, $0 \le h_i \le 1$. The sum of the leverages $h_1 + \cdots + h_k = Q$, where Q is the total number of predictors including the intercept (that is $Q = p + 1$ when there is an intercept in the model and $Q = p$ if there is no intercept). Thus the average value of the leverage is Q/k, and estimates of regression coefficients are most efficient when all the leverage values are close to Q/k.

If $h_i = 0$, this implies that the fitted (predicted) value of the effect size for the i^{th} study would be the same even if that study were not part of the data used to estimate the regression coefficients. In one sense, this implies minimal influence. If $h_i = 1$, this implies that the fitted value of the i^{th} study could not be estimated without the data from that study, in other words, the fitted value of that study depends entirely on data from that study. This latter situation is equivalent to saying that there is a regression coefficient (or linear combination of regression coefficients) whose estimate is determined entirely by the data from i^{th} study. In other regression contexts, reference value of $2q/k$ as been suggested as indicating a study of high leverage.

The term leverage arises from a mechanical analogy. Imagine a scatterplot of the effect size versus one predictor. In this one predictor situation, the studies that have *x* (predictor) values that are far from the center of the data will have high leverage because moving them up or down would have large influence of the regression slope. When there is more than one predictor, there may be studies whose combination of predictor values is far from the center in a multivariate sense. The leverage diagnostic may reveal such multivariate outliers that are not obvious from looking at predictors one at a time.

Cook's Distance

Cook's distance, D_i, for the i^{th} study is a measure of how much the estimated regression coefficients change (on the average) when the i^{th} study is deleted from the dataset. Like the studentized and jackknifed residuals, D_i is standardized, but unlike them it is in a squared (distance-squared) metric.

One can think of D_i as the squared difference between **b** the vector of regression coefficient estimates estimated from all studies and the vector $\mathbf{b}_{(i)}$ of regression coefficients estimated from all studies except the i^{th} study, divided by the variance of **b** that is

$$D_i = \frac{\left(\mathbf{b} - \mathbf{b}_{(i)}\right)' \mathbf{V}^{-1} \left(\mathbf{b} - \mathbf{b}_{(i)}\right)}{p+1} = \frac{w_i h_i e_i^2}{q\left(1-h_i\right)^2} \ , \tag{1.23}$$

where **V** is the covariance matrix of **b**. In other regression contexts, the value $4/(k - Q)$ been suggested to help identify studies that have large influence.

***DF*ITTS**

*DF*ITTS is a diagnostic that describes the change in the fitted (predicted) value of the i^{th} study that would arise as a consequence of deleting the i^{th} study from the data to compute the regression coefficients used to compute the fitted value. *DF*FITS is defined as

$$DFFITS_i = \hat{T}_i - \hat{T}_{(i)i} = e_i \sqrt{\frac{w_i h_i}{\left(1-h_i\right)^2}} \ , \tag{1.24}$$

where

$$\hat{T}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} \tag{1.25}$$

and

$$\hat{T}_{(i)i} = b_{(i)0} + b_{(i)1} x_{i1} + \cdots + b_{(i)p} x_{ip} \ , \tag{1.26}$$

where $b_{(i)0}$, …, $b_{(i)p}$ are the regression coefficients estimated with the i^{th} study removed from the dataset.

Like the jackknifed residual, *DF*FITS is designed to better reveal cases where the i^{th} study does not fit the same model as the other studies. By removing the (potentially distorting impact of the i^{th} study from the computation of the regression coefficients, the jackknifed residual sometimes makes it easier to see how different an observed effect size is from what is expected if that study fit the meta-regression model that is appropriate for all of the other studies. In other regression contexts, the reference value $2\sqrt{q/n}$ has been suggested for identifying studies with potentially large impact on fitted values.

Variance

The variance v_i of the i^{th} study is the conditional (estimation error) variance of the effect size in the i^{th} study. Because v_i depends on the sample size in each study, v_i can vary substantially across studies.

Tau Squared

Tau-squared, τ^2, is the estimate of between-studies variance among effect size parameters at any point on the prediction line. An assumption of the meta-regression is that the true variance of effect sizes is the same for all values of the covariate.

Sum

Sum is the total variance of the i^{th} effect size, which is v_i in fixed effect meta-regression or $\tau^2 + v_i$ in random effects meta-regression.

Weight

The weight of the i^{th} study, w_i, is the actual (raw) weight assigned to this study in the analysis, namely the reciprocal of the total variance, namely $w_i = 1/v_i$ in fixed effects meta-regression and $w_i = 1/(\tau^2 + v_i)$ in random effects meta-regression.

Percent Weight

Percent weight for the i^{th} study is the percentage of the total weight accorded to study, that is w_i divided by the sum of all study weights.

Variance Inflation Factor

This diagnostic is not located on the diagnostic screen. Rather, click Computational options > Variance inflation factor and the this will be displayed on the main analysis screen adjacent to each covariate.

The variance inflation factor VIF_j for the j^{th} covariate is a diagnostic designed to provide information about the colinearity of the covariate set. One of the consequences of colinearity is that it increases the variance (the square of the standard error of) the regression coefficient estimates. If the standard error of a regression coefficient estimate is too large, it may be difficult to meaningfully interpret that estimate. For example, suppose that a particular coefficient expresses the difference between the average of two groups of standardized mean difference effect sizes, that the coefficient estimate is 1.0 and the standard error of that coefficient is 4. In such a case, it is difficult to draw an informative conclusion because the results imply a 95 percent confidence interval for the mean difference between group mean effects of -3 to +5, a range which is consistent with substantially different substantive conclusions.

The VIF_j indicates how much greater the variance of the regression coefficient estimate b_j for the j^{th} covariate is than it would have been if the covariates were totally uncorrelated. A *VIF* value of 4 for a

particular covariate indicates that the standard error of that regression coefficient is twice as large as it would have been if the covariate were uncorrelated with all the other covariates.

A high VIF_j value for the j^{th} covariate does not necessarily mean that the standard error of the coefficient of that covariate is too large for the estimate to be meaningful. For example, suppose that a particular coefficient expresses the difference between the average of two groups of standardized mean difference effect sizes, that the coefficient estimate is 1.0, and the standard error of that coefficient is 0.1 In such a case, one can still draw an informative conclusion even if $VIF = 4$ because the results imply a 95 percent confidence interval for the mean difference between group mean effects of 0.8 to 1.2, a range which is consistent with substantially the same substantive conclusion of a very large difference between the group mean effect sizes.

Note that VIF_j is not a property of the j^{th} covariate alone, but depends on all of the other covariates as well. Therefore removing one covariate from the covariate set may change (sometimes drastically) the *VIF* values of several other covariates.

COVARIANCE

To navigate to this page click More results > Covariance matrix

ADD OTHER VARS

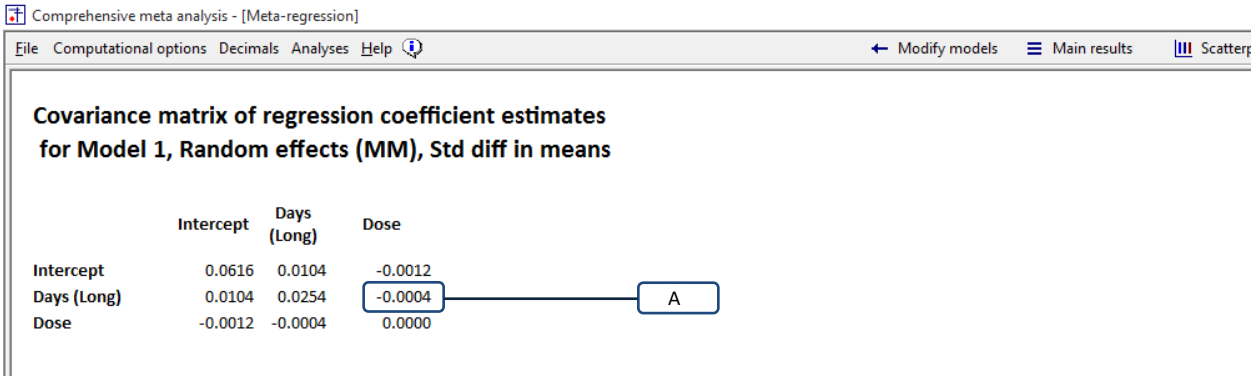


Figure 32 | Covariance matrix

This page gives us the covariance matrix of the *B* values.

To understand what these covariances represent, imagine that we draw a sample of studies, run the regression, and get an estimate of B_{DAYS} and B_{DOSE}. We repeat this process *j* times, and each time get an estimate of B_{DAYS} and B_{DOSE}. Then, we compute the covariance of B_{DAYS} and B_{DOSE} over the *j* samples. This covariance would be −0.0004 [A]

The same idea applies to all cells in the matrix.

CORRELATIONS

To navigate to this page click More results > Correlation matrix

ADD OTHER VARS

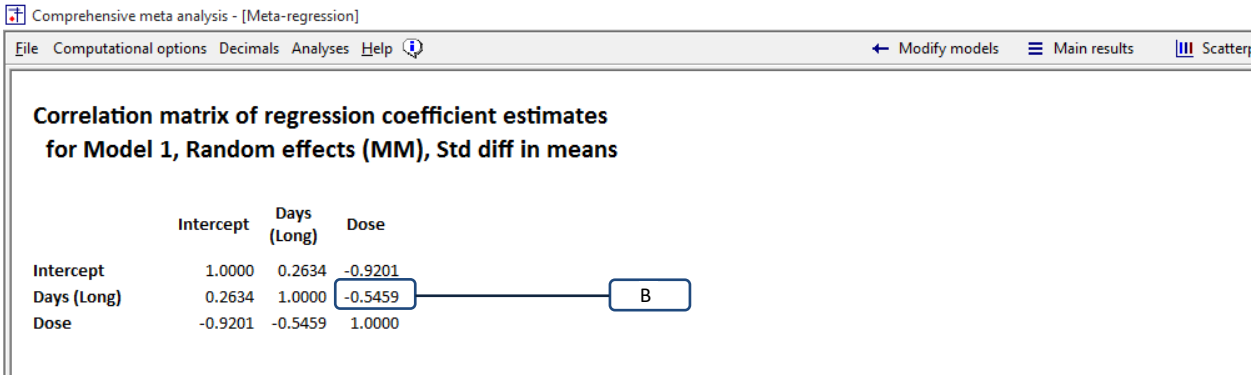


Figure 33 | Correlation matrix

This page gives us the correlation matrix of the *B* values.

To understand what these covariances represent, imagine that we draw a sample of studies, run the regression, and get an estimate of B_{DAYS} and B_{DOSE}. We repeat this process *j* times, and each time get an estimate of B_{DAYS} and B_{DOSE}. Then, we compute the covariance of B_{DAYS} and B_{DOSE} over the *j* samples. This correlation would be −0.5459 [B]

The same idea applies to all cells in the matrix.

When the correlation between two covariates is high (close to 1.0 or close to −1.0), this tells us that the two are highly confounded, and it is therefore difficult to isolate the unique impact of each.

ASSESSING CHANGE IN THE MODEL

Suppose we run a model with only *Dose* as a covariate. Then we run a model with *Dose* plus *SUD: Y*. Just as we can report statistics for either model as compared with the null model (the intercept only) we can also report statistics for the second model as compared with the first.

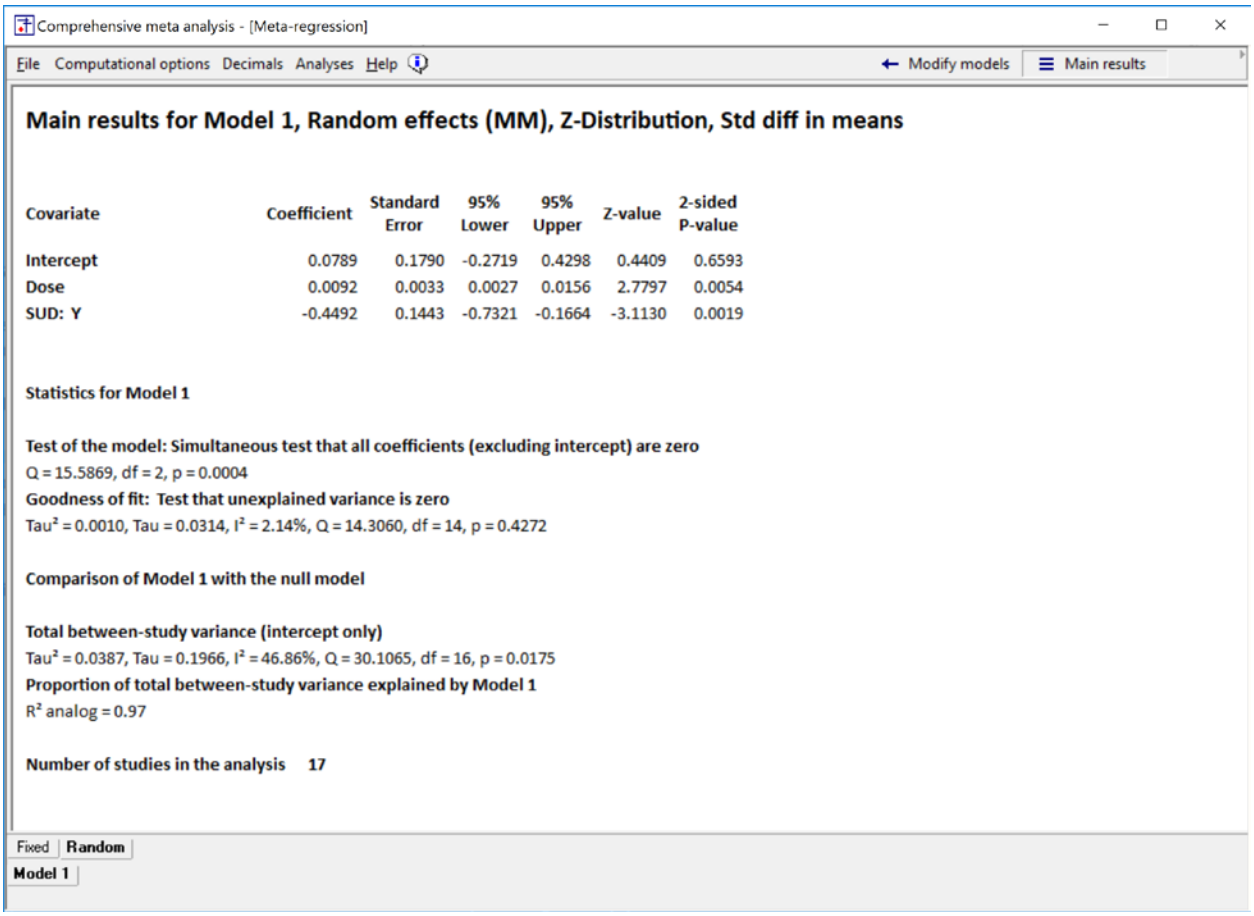


Figure 34 | Main results | Random-effects

For example, consider the [Main results] screen shown in Figure 34. On this screen, the test of the model, goodness of fit, estimates of T^2 and R^2, apply to the full model (*Dose* and *SUD: Y*).

Suppose we want to know these statistics if (a) we include only Dose, and then (b) we include Dose and SUD: Y. To get this information we would need to run two analyses, adding one covariate at each step. We can do this manually, but the program will also do it automatically. Additionally, the program is able to collate the results from each step into a single table.

To make it clear how the increments work, we will the run series of analysis both manually and then automatically, and then show how the two compare.

In Figure 35 we include a tick-mark for the intercept only.

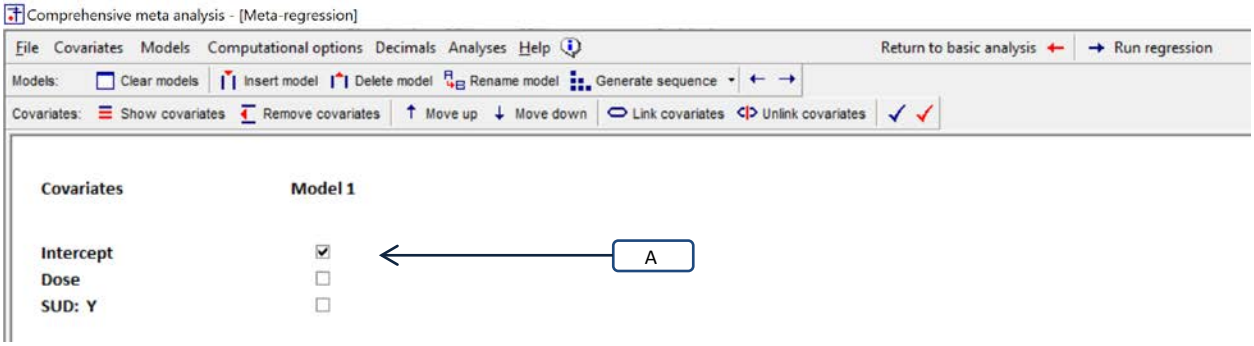


Figure 35 | Setup | Intercept only

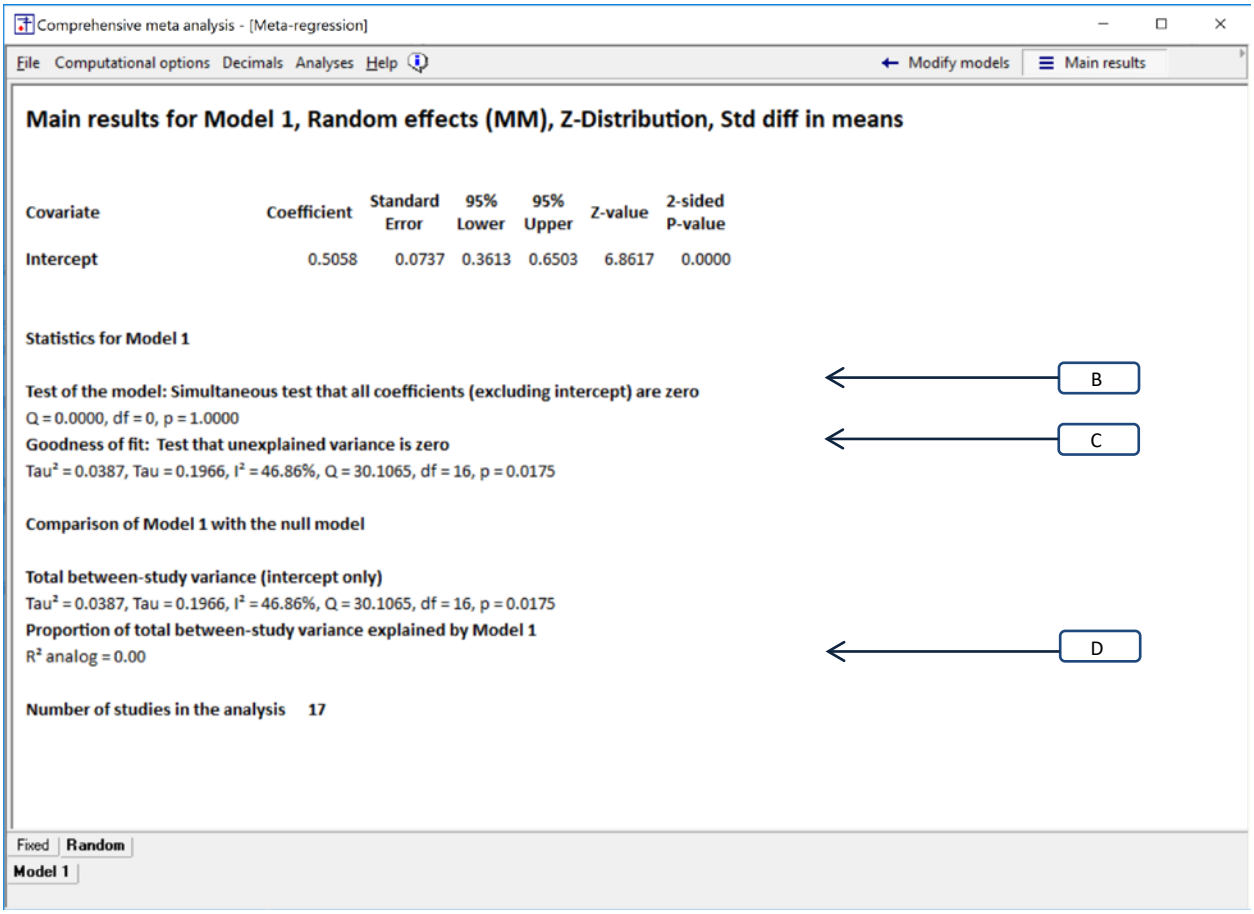


Figure 36 | Main results | Intercept only

With only the intercept in the model (Figure 36),

- Explained by the model $Q = 0.0000$, $df = 0$, $p = 1.0000$ [B]
- Unexplained by the model $T^2=0.0387$, $Q = 30.1065$, $df = 16$, $p= 0.0175$ [C]
- R^2 for the model is 0.00 [D]

In Figure 37 we add a tick-mark for Dose [A] and re-run the analysis.

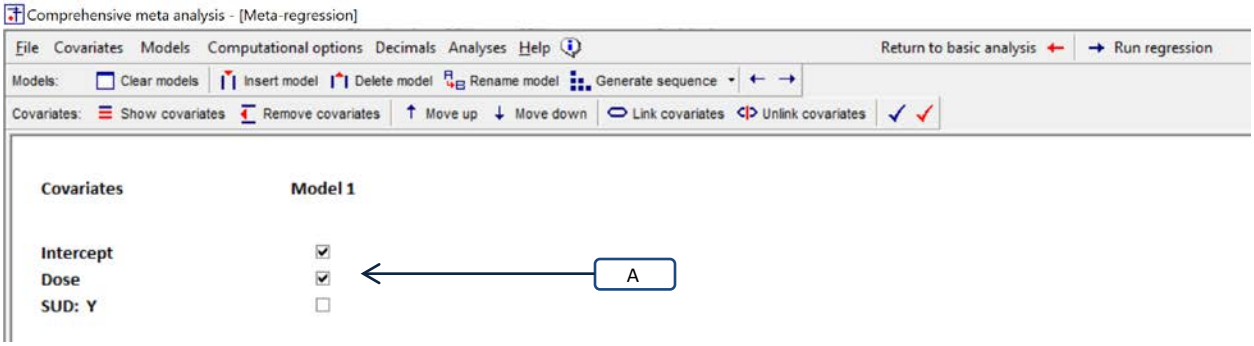


Figure 37 | Setup | Intercept + Formulation

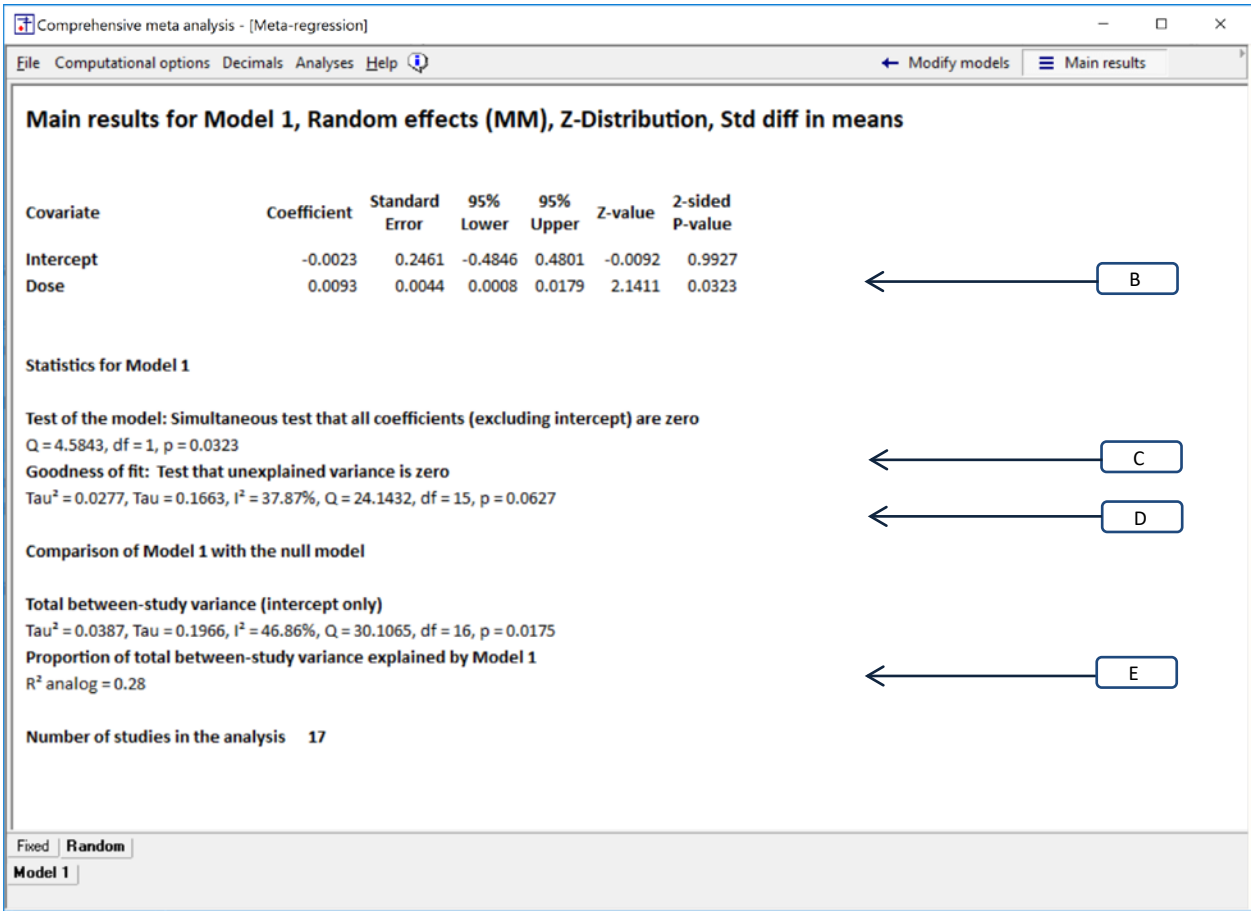


Figure 38

With intercept + Dose in the model Figure 38

- Impact of Dose Z=2.1411, p=0.0323 [B]
- Explained by the model Q = 4.5843, df=1,p=0.0323 [C]
- Unexplained by the model T2=0.0277, Q = 24.1432, df = 15, p = 0.0627 [D]
- R^2 for the model is 0.2847 [E]

In Figure 39 we add a tick-mark for SUD: Y and re-run the analysis [A]

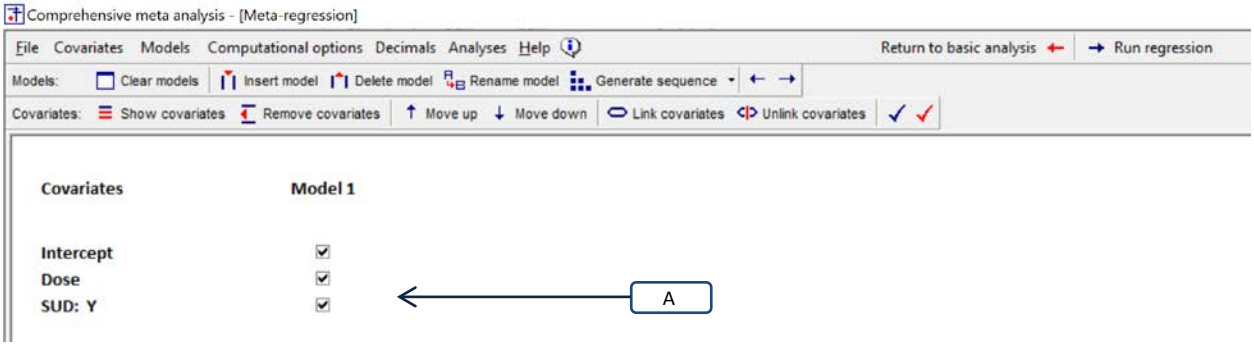


Figure 39

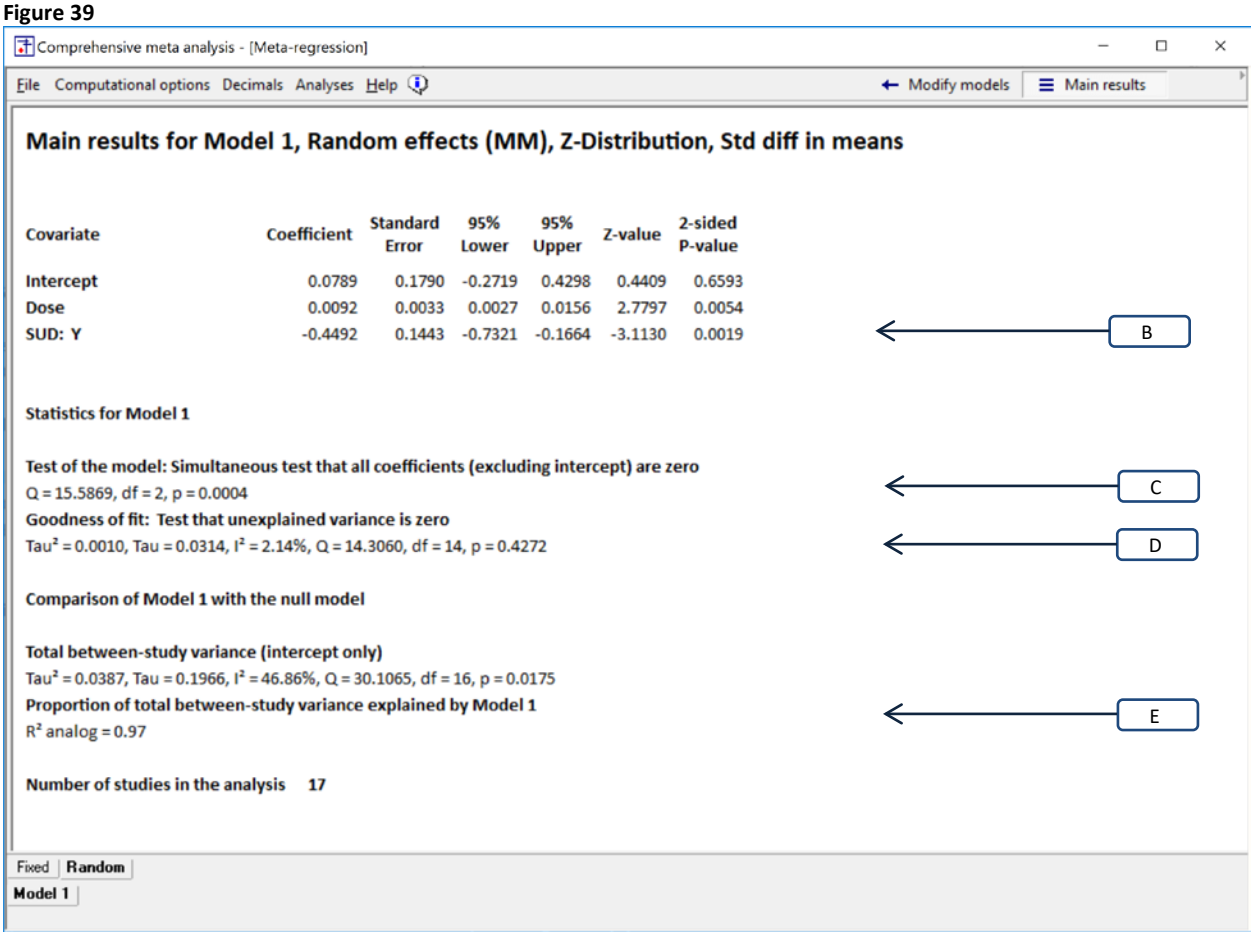


Figure 40

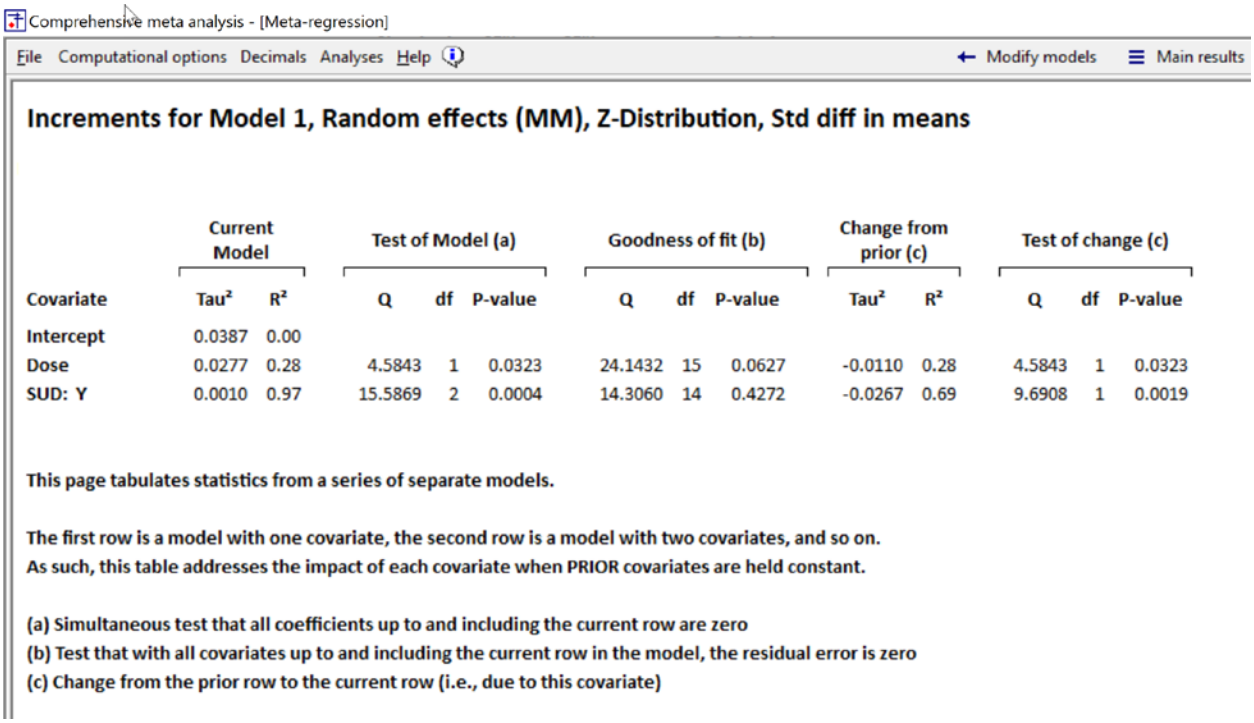
With Intercept + Dose + SUD:Y in the model Figure 40

- Unique impact of SUD: Y Z=−3.1130, p=0.0019 [B]
- Explained by the model Q = 15.5869, df= 2, p=0.0004 [C]
- Unexplained by the model T2=0.0010, Q = 14.3060, df = 14, p = 0.4272 [D]
- R^2 for the model is 0.97 [E]

Alternatively, we could have run the full model only, and gone to the increments page.

To navigate to this page —

- Run the analysis that includes all the covariates
- Click More results > Increments
- Select the statistical model tab (Fixed or random)



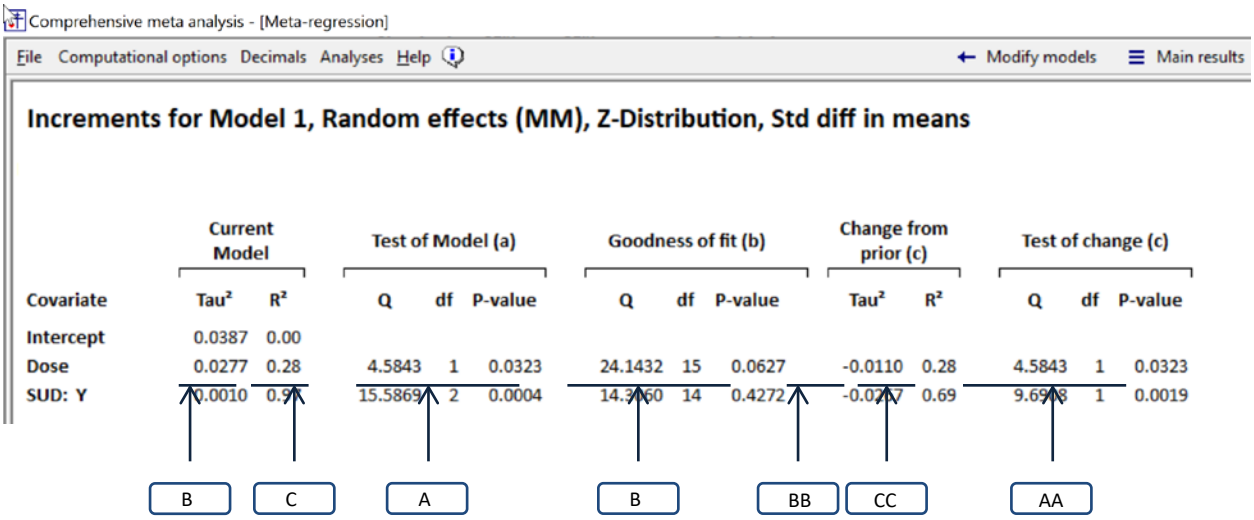


Figure 41 | Main results | Intercept + Allocation + Year + Dose

Every row in this table copies information from a separate analysis.

- The row labeled Intercept copies information from Figure 36. The model is Intercept alone.
- The row labeled Dose copies information from **Error! Reference source not found.**. The model is Intercept + Dose.
- The row labeled SUD: Y copies information from **Error! Reference source not found.**. The model is Intercept + Dose +SUD: Y

Every column in this table corresponds to a section in the prior figures

- [B] Current model T^2 corresponds to section [B] in prior screens
- [C] Current model R^2 [C] corresponds to section [C] in prior screen
- [A] Test of model [A] corresponds to section [A] in prior screens
- [B] Goodness of fit [B] corresponds to section [B] in prior screens

Additionally, this table presents information about change from one row to the next

- [BB] Change in T^2 from prior row
- [CC] Change in R^2 from prior row
- [AA] Test for change in T^2 from prior row

Suppose we want information about the model that includes Dose only. On the line labeled Dose we see that T^2 is 0.0277, R^2 is 0.28, and the model is statistically significant (Q = 4.5843, df = 1, p = 0.0323). These statistics were copied from the analysis in **Error! Reference source not found.**.

Similarly, suppose we want information about the model that includes Dose and SUD: Y. On the line labeled SUD: Y we see that T^2 is 0.0010, R^2 is 0.97, and the model is statistically significant (Q = 15.5869, df = 2, p = 0.0004). These statistics were copied from the analysis in **Error! Reference source not found.**.

The columns labelled [AA], [BB] and [CC] are unique to this screen, and address *the change* as we move from one model to the next. The column labeled "Change from prior" gives the change in T^2 and in R^2. The column labeled "Test of change" is the corresponding test of statistical significance.

The column labeled Test of change in Figure 42 corresponds to the impact of each covariate *at the point that it is entered into the model*. Thus,

The change for Dose corresponds to the impact of Dose with no covariates held constant. The p-value of 0.0323 in Figure 42 corresponds to the p-value of 0.0323 in Figure 38

The change for SUD: Y corresponds to the impact of SUD: Y with Dose held constant. The p-value of ____ in ____ corresponds to the p-value of ____ in _____.

Note. The tests in the earlier screens report Z rather than Q. Where Z is in linear units, Q is in squared units. Therefore, we would square the Z-value in the earlier figures to get the Q-value in this figure.

NEED TO EXPLAIN THAT ROWS SHOW UNIQUE IMPACT OF EACH COVARIATE WHILE MODEL SHOWS COMBINED IMPACT OF ALL

SOME FUNDAMENTAL ISSUES IN META-REGRESSION

FIXED-EFFECT VS. RANDOM-EFFECTS

Most meta-analyses are based on either of two statistical models – the fixed-effect model or the random-effects model. In this chapter we will explain what the models mean, how the selection of a model provides the framework for the analysis, and how to select a model.

We will start with a simple analysis, where we are working with one set of studies, to establish the basic principles. Then, we will show how the same ideas apply to an analysis with two or more subgroups of studies. Finally, we will show how they apply in the case of a meta-regression.

A simple analysis

Consider a simple meta-analysis, where we have one set of studies, and no moderators. Our goal is to compute the summary effect size based on this set of studies.

- The fixed-effect model applies if all studies are drawn from a single population. The studies share a common *true* effect-size, and so the *true* effect size is *fixed*, or constant.

- The random-effects model applies if the studies are drawn from a universe of populations. The *true* effect size varies from one population to the next, and the studies are sampled at *random* from this universe.

Note that we are using the term "population" to refer not only to the subjects, but also to the methods, time-frame, assessments, and any factor that could have an impact on the true effect size. On this basis, the fixed-effect model would apply only if the studies are virtual replicates of each other. By contrast, if we identify studies in the literature that were performed by different people, working with different populations, the random-effects model would apply. While these studies may be addressing the same fundamental question, the true effect size probably varies from study to study.

The selection of a statistical model is critically important for several reasons.

First, the statistical model determines how we can generalize the results

- Under the fixed-effect model all the studies in our analysis come from one population. The results apply only to that specific population.

- Under the random-effects model the populations in our analysis have been sampled from a defined universe of populations. The results apply to that universe.

Second, it sets up a framework for the analysis, establishing what questions we can ask and how to interpret the results.

- Under the fixed-effect model we assume that all studies share a common *true* effect size, and our goal is to estimate *this common* parameter. We don't estimate the variance in true effects, since (by definition) the intervention has precisely the same effect in all studies.

- Under the random-effects model we expect that the *true* effect size varies from study to study, and our goal is to estimate *the mean* of these parameters. We also estimate the variance in true effects. Indeed, this may be a key part of the analysis.

The selection of a model affects how weights are assigned to the studies. This affects the precision of the summary estimate, and it also affects the summary estimate itself.

- Under the fixed-effect model there is only one level of sampling (the subjects in each study are sampled from all subjects in the study's population) and therefore only one source of sampling error (the observed effect size in each study differs from the true effect size for the common population). The error variance for study (i) is quantified as V_i, and the weight assigned to each study is the inverse of this variance, or $1/V_i$.

- Under the random-effects model there are two levels of sampling (the subjects in each study are sampled from all subjects in the study's population, and the study populations are sampled from the universe of study populations) and therefore two sources of sampling error (the observed effect in each study differs from the true effect for that study's population, and the mean true effect for the sampled populations differs from the mean for the universe of populations). The first error variance is quantified as V and the second as T^2. The total error variance for each study is then V_i+T^2, and the weight assigned to each study is the inverse of this variance, or $1/(V_i+T^2)$

The difference in the weights between the two models affects the precision with which we estimate the summary effect. In the theoretical case where the within-study variance is the same for all studies, the different weights described above would have the following impact on the standard error of the summary effect.

Under the fixed-effect model the standard error of the summary effect is

$$SE_M = \sqrt{\frac{V}{N}} \tag{1.27}$$

where V is the (common) within-study variance and N is the number of subjects accumulated across studies. By contrast, under the random-effects model, the standard error of the summary effect is

$$SE_M = \sqrt{\frac{V}{N} + \frac{T^2}{k}} \tag{1.28}$$

where T^2 is the between-study variance and k is the number of studies. The addition of the second term under the radical addresses the additional sampling error that comes with the fact that we are sampling studies from a universe of different populations, rather than limiting ourselves to one population. The inclusion of this term widens the confidence interval, and allows us to generalize to this wider universe.

The difference in weights also affects the estimate of the summary effect itself. Under both models, studies with more precise estimates of the effect size (typically the larger studies) are given more weight

than studies with less precise estimates. However, this difference is more pronounced under the fixed-effect model and more moderate under the random-effects model.

Concretely, under the fixed-effect model the weight assigned to each study is

$$W_i = \frac{1}{V_i} \tag{1.29}$$

Where V_i is the variance in study (i). If V_1 is 0.01 and V_2 is 0.02, then

$$W_1 = \frac{1}{V_1} = \frac{1}{.01} = \frac{1}{.01} = 100.0 \tag{1.30}$$

$$W_2 = \frac{1}{V_2} = \frac{1}{.02} = 50.0 \tag{1.31}$$

Study 1 will be given two times as much weight as Study 2. By contrast, under the random-effects model, the weight assigned to each study is

$$W_i = \frac{1}{V_i + T^2} \tag{1.32}$$

Where T^2 is the between-study variance. Since T^2 is the same for all studies, the inclusion of this term serves to limit the difference in the weights across studies. For example, if T^2 is .02, then

$$W_1 = \frac{1}{V_1 + T^2} = \frac{1}{.01 + .02} = \frac{1}{.03} = 33.3 \tag{1.33}$$

$$W_2 = \frac{1}{V_2 + T^2} = \frac{1}{.02 + .02} = \frac{1}{.04} = 25.0 \tag{1.34}$$

In this case the larger study gets 32% more weight than the smaller one.

The idea that the larger studies tend to be more dominant under the fixed-effect model and less dominant under the random-effects model follows from the logic of the two sampling frames. Under the fixed-effect model all studies are estimating the same value. So, if one study has a very precise estimate of that value while another has a very poor estimate of that value, we will give substantially more weight to the first study. By contrast, under the random-effects model each study is estimating the effect size in a different population, and we use these populations to estimate the mean effect across all populations in the relevant universe of populations. If we know the effect size in one study precisely we know the effect for that population but we don't know if that population falls near the mean of all populations. Therefore, we don't want it to dominate the analysis and we apply a more moderate weight.

How do we choose a model?

The selection of a statistical model is determined by our understanding of the sampling frame. If all studies are sampled from the same population and are identical to each other in all relevant respects, then the fixed-effect model applies. Otherwise, the random-effects model applies. Following are two fictional examples.

Fixed-effect sampling frame. Suppose that a drug company is running a series of studies. In each study patients are randomized to either drug or placebo, and then assessed on some measure. All samples are drawn from the same population, and all studies are identical to each other in all respects – they employ the same researchers, time frame, dose, and so on. It follows that all studies are estimating the same parameter, which is the impact of the drug in this specific population, under these specific conditions. Put another way, if the true effect size for the entire population is 0.50, then the observed effect size in any of these studies would approach 0.50 as the number of subjects in that study approached the number of people in the population. For this reason, the fixed-effect model applies.

Random-effects sampling frame. Suppose that a drug company searches the literature and locates ten studies that assessed the impact of their drug. In each study patients were randomized to either drug or placebo, and then assessed on some measure after a short period of time. Each study sampled patients from a different hospital, and so the populations varied in age, health, and various other dimensions. Experience tells us that the impact of any intervention will vary (at least a little) from one population to the next, and so it follows that each study is estimating a different parameter than the others. Put another way, if we increased the sample size in each study to include the full population at that study's hospital, it would *not* be the case that the observed effect size in all the studies would converge on the same value. Rather, the effects would vary about some mean. For this reason, the random-effects model applies.

The model provides context for the forest plot.

The results of *both* these meta-analyses are shown in Figure 42. The data for the individual studies is the same under both models, but the analysis will depend on whether we apply the fixed-effect model or the random-effects model. We deliberately set up the (fictional) data this way to emphasize the fact that the selection of a model depends on our understanding of the sampling frame, and not on the dispersion observed in the analysis. If the studies are sampled from one population (and are essentially replicates of each other) we choose the fixed-effect model. If the studies are sampled from multiple populations, we choose the random-effects model. If we don't know which of these applies, then we are simply playing with numbers, and have no business doing the analysis.

Suppose that the studies all come from the same population, as per the first example above. All of the studies are estimating the true effect size in the same population. The observed effect size varies from study to study, but that variance is due entirely to random sampling error. If we could somehow remove the sampling error (and plot the true effects) all the effects would be exactly the same. We report that the common effect size in this specific population, and we stop there. If we are talking about one population, we are talking about one effect size. And one effect size cannot vary – it's one number.

Alternatively, suppose that the studies come from a sample of populations, and we are using this sample to generalize to a universe of populations, as per the second example above. Each study is estimating the true effect size in one unique population. The observed effect size varies from study to study. Some of that variance is due to random sampling error, but some reflects the fact that the intervention has more of an impact in some populations than in others. If we could somehow remove the sampling error

(and plot the true effects), the effects would vary from one population to the next. We report the mean effect size. And we also report how the effect size varies across the populations from which the populations in the analysis were sampled.

Since the two models reflect different assumptions about the sampling process, and have different goals, there is generally no reason to compare the results of the two. Here, purely as an educational exercise, we show how the statistics differ by the statistical model. Our goal is twofold. First, we want to show how the formulas reflect the sampling frame and the goals outlined above. Second, we want to show what happens if someone should be using the random-effects model but uses the fixed-effect model instead.

When the fixed-effect model is appropriate we are estimating the mean for the one population of interest and so the error term is relatively small. When the random-effects model is appropriate we are estimating the mean for the populations in the analysis, and then using these to generalize to all populations in the universe. Therefore, the error term is larger.

In the drug-company example, the mean effect size *in this population* is 0.50 with a standard error of ____ and a confidence interval of _____. In the example where studies are pulled from the literature, the mean effect size *in the universe of populations* from which these studies were sampled is 0.50 with a standard error of ____ and a confidence interval of _____.

To show the impact of the model on the error variance, we can use an example where we are trying to estimate the mean in one group. We will also assume that the population variance (V) is the same in all studies. Under the fixed-effect model the error variance of the mean will be V/N where N is accumulated across studies. Under the random-effects model the error variance will be V/N + T2/K where k is the number of studies.

The term V/N refers to within-study variance, and is the same in both cases. It reflects the fact that the mean for each sample is not the same as the mean in the corresponding population. This error variance tends to diminish as N increases, regardless of whether the N is located in one study or distributed across many studies. The second term T2/k reflects the between-study variance. It reflects the fact that the mean for these populations (the ones in the analysis) is not the same as the mean for all populations in the universe. This error variance tends to diminish as K increases. Critically, the two components of variance are additive and independent of each other. Increasing N has zero impact on the second element of the variance. If T2 is substantial and k is small, then the error variance will be non-trivial.

, and where the population variance is the same in all studies. The e

If the population variance is V, then the error variance of the mean will be V/n, where n is the sample size in that study. This reflects the fact that as the sample size increases, values on opposite sides of the mean will tend to cancel each other out. The same idea can be extended to a meta-analysis. If the population variance in each study is V, then the error variance of the mean will be V/N, where N is the total sample size, accumulated across studies. Since the impact of N is for values on either side of the mean to balance each other, this works across studies as well as within studies.

is 0.50 with a standard error of 0.033. We use these values to construct a confidence interval of 0.435 to 0.565. We also use these values to test the null hypothesis that the common effect size is 0.000. The test yields a Z-value of 15.000 and a p-value of < 0.001. These studies are all sampled from one population, and so our goal is to describe the effect in this population – we cannot generalize to any other populations. We don't estimate the variance in true effects. All the studies are estimating a common parameter, and a single value cannot vary.

Alternatively, suppose that the studies are sampled from different populations, as per the second example. We would conclude that the mean effect size is 0.50 with a standard error of 0.046. We could use these values to construct a confidence interval of 0.411 to 0.589. We could also use these values to test the null hypothesis that the common effect size is 0.000. The test yields a Z-value of 10.954 and a p-value of < 0.001.

These studies have been sampled from a universe of populations, and so our goal is to generalize to this universe. Critically, we want to estimate not only the mean effect size, but also how the true effects vary across populations. The standard deviation of true effects (T) is 0.094. A back-of-the-envelope calculation tells is that the effect size will vary from 0.313 in some populations to 0.687 in others.

The analysis is shown in Figure 42. The common effect size is 0.500 with a standard error of 0.033.

The confidence interval is computed using

$$\begin{aligned} LL_d &= d - 1.96 \times SE_d \\ UL_d &= d + 1.96 \times SE_d \end{aligned} \tag{1.35}$$

Which here is

$$\begin{aligned} LL_d &= 0.50 - 1.96 \times 0.033 = 0.435 \\ UL_d &= 0.20 + 1.96 \times 0.033 = 0.565 \end{aligned} \tag{1.36}$$

The confidence interval is 0.435 to 0.565, which tells us that the true effect size in this population probably falls somewhere in this range.

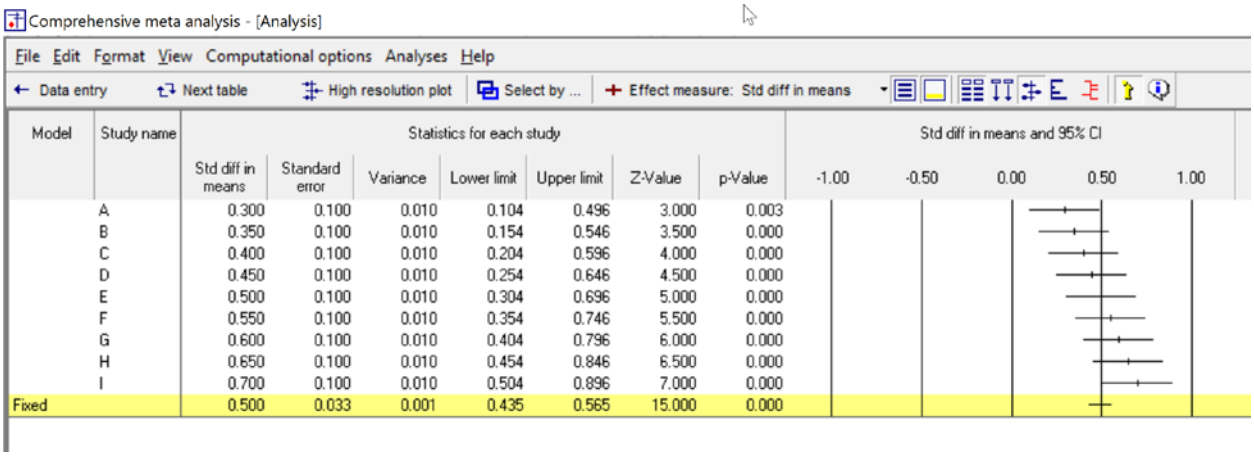
Similarly, to test the null hypothesis that the true effect size is zero we could compute a Z-value using

$$Z = \frac{d}{SE_d} \ . \tag{1.37}$$

Here,

$$Z = \frac{0.500}{0.033} = 15.000 \tag{1.38}$$

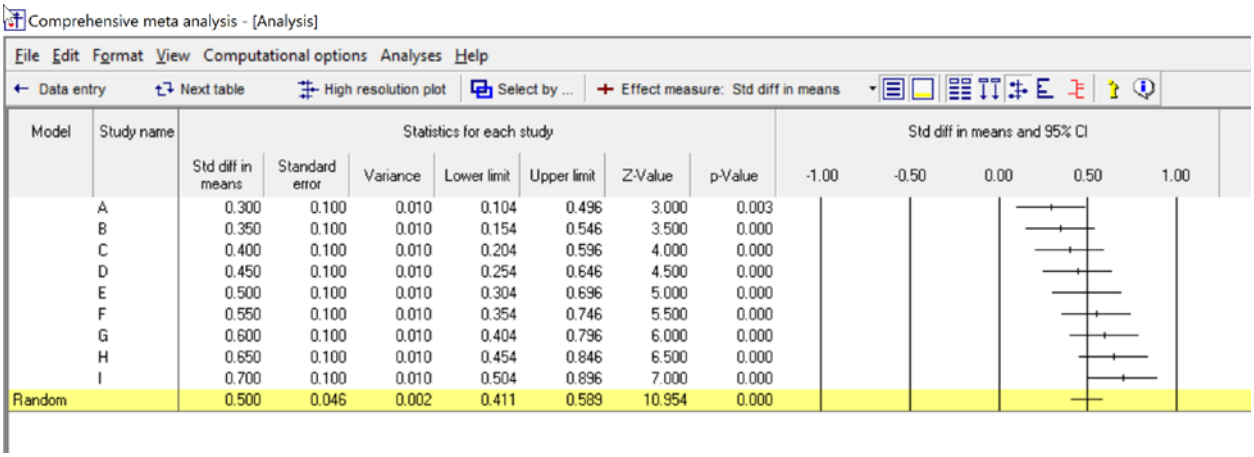
with a corresponding p-value of < 0.001. We reject the null hypothesis and conclude that, in this population, the drug increases the score as compared with placebo.



Model	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	A	0.300	0.100	0.010	0.104	0.496	3.000	0.003
	B	0.350	0.100	0.010	0.154	0.546	3.500	0.000
	C	0.400	0.100	0.010	0.204	0.596	4.000	0.000
	D	0.450	0.100	0.010	0.254	0.646	4.500	0.000
	E	0.500	0.100	0.010	0.304	0.696	5.000	0.000
	F	0.550	0.100	0.010	0.354	0.746	5.500	0.000
	G	0.600	0.100	0.010	0.404	0.796	6.000	0.000
	H	0.650	0.100	0.010	0.454	0.846	6.500	0.000
	I	0.700	0.100	0.010	0.504	0.896	7.000	0.000
Fixed		0.500	0.033	0.001	0.435	0.565	15.000	0.000

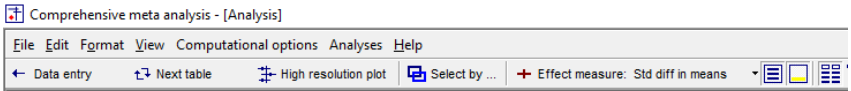
Figure 42

We say nothing about variation in effect size across studies. All studies in this analysis are estimating the same parameter, which is the impact of the drug in this one population under a specific set of conditions. It makes no sense to ask *how* this parameter varies – by definition, it *cannot* vary, since it's only one value. Rather, all variation in the observed effects must reflect sampling error alone. Put another way, if the true effect size in this population is 0.50, then if every study had an extremely large sample size (and only trivial sampling error) all the effects in the forest plot would converge on this common effect size.

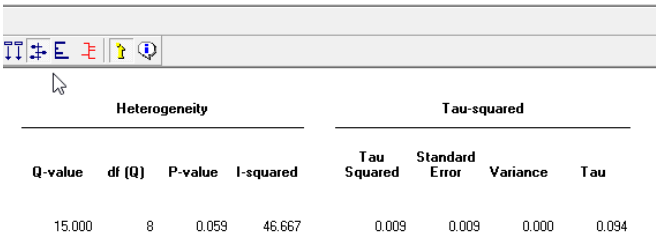


Model	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	A	0.300	0.100	0.010	0.104	0.496	3.000	0.003
	B	0.350	0.100	0.010	0.154	0.546	3.500	0.000
	C	0.400	0.100	0.010	0.204	0.596	4.000	0.000
	D	0.450	0.100	0.010	0.254	0.646	4.500	0.000
	E	0.500	0.100	0.010	0.304	0.696	5.000	0.000
	F	0.550	0.100	0.010	0.354	0.746	5.500	0.000
	G	0.600	0.100	0.010	0.404	0.796	6.000	0.000
	H	0.650	0.100	0.010	0.454	0.846	6.500	0.000
	I	0.700	0.100	0.010	0.504	0.896	7.000	0.000
Random		0.500	0.046	0.002	0.411	0.589	10.954	0.000

Figure 43



Model		Effect size and 95% confidence interval					Test of null (2-Tail)	
Model	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value
Fixed	9	0.500	0.033	0.001	0.435	0.565	15.000	0.000
Random	9	0.500	0.046	0.002	0.411	0.589	10.954	0.000



Heterogeneity				Tau-squared			
Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
15.000	8	0.059	46.667	0.009	0.009	0.000	0.094

Figure 44

The analysis is shown in Figure 42. The mean effect size is 0.500 with a standard error of 0.046.

The confidence interval is computed using

$$\begin{aligned} LL_d &= d - 1.96 \times SE_d \\ UL_d &= d + 1.96 \times SE_d \end{aligned} \tag{1.39}$$

Which here is

$$\begin{aligned} LL_d &= 0.50 - 1.96 \times 0.046 = 0.411 \\ UL_d &= 0.50 + 1.96 \times 0.046 = 0.589 \end{aligned} \tag{1.40}$$

The confidence interval is 0.411 to 0.589, which tells us that the mean effect size in the universe of relevant populations probably falls somewhere in this range.

Similarly, to test the null hypothesis that the true effect size is zero we could compute a Z-value using

$$Z = \frac{d}{SE_d} \ . \tag{1.41}$$

Here,

$$Z = \frac{0.500}{0.046} = 10.954 \tag{1.42}$$

with a corresponding p-value of < 0.001. We reject the null hypothesis and conclude that, in the universe of relevant populations, the drug increases the score on the test as compared with placebo.

We would also estimate the variation in effect size across studies. The prediction interval (PI) gives us the range of effect sizes that includes some 95% of all relevant populations, and is given by

$$\begin{aligned} PI_{LL} &= d - 2T \\ PI_{UL} &= d + 2T \end{aligned} \tag{1.43}$$

Here, the standard deviation of true effects is estimated as 0.0935. In the universe of populations from which these studies were sampled, most studies that employed 10 mg, will have a true effect size in the range of

$$\begin{aligned} PI_{LL} &= d - 2T = 0.50 - 2 \times 0.0935 = 0.313 \\ PI_{UL} &= d + 2T = 0.50 + 2 \times 0.0935 = 0.687 \end{aligned} \tag{1.44}$$

(This formula assumes that the mean and standard deviation are known precisely. Formulas that allow for uncertainty in these estimates are presented elsewhere in this volume.)

Subgroups

Consider a subgroups analysis, where we have two (or more) sets of studies. Our goal is to compare the mean effect size across subgroups.

- The fixed-effect model applies if all studies within each subgroup are drawn from a single population. The studies within each subgroup share a common *true* effect-size, and so the *true* effect size is *fixed*, or constant.

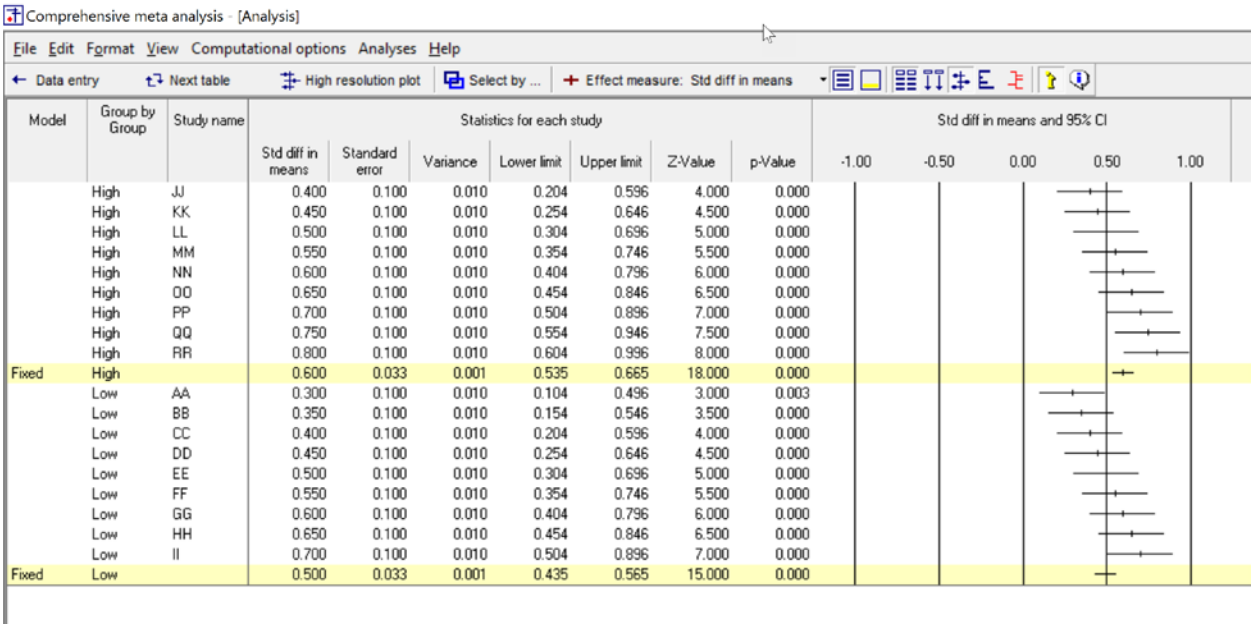
- The random-effects model applies if the studies within each subgroup are drawn from a universe of populations. The *true* effect size within each subgroup varies from one population to the next, and the studies are sampled at *random* from this universe.

All of the issues discussed above for a simple analysis, apply here as well. That is,

Subgroups

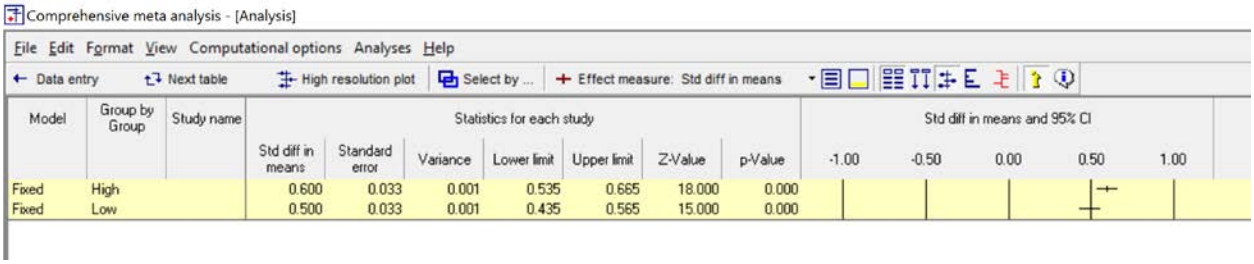
Consider the same example as above, where all studies are drawn from the same population, and in each study patients are randomized to either the drug or a placebo. In this case, however, the dose of the drug is set (by random assignment) to 10 mg. in some studies, and to 20 mg. in the other studies.

The studies which employ a 10 mg. dose are all estimating the same parameter, which is the impact of this dose in this specific population. If the true effect size for this dose is 0.50, then the observed effect size in any of these studies would approach 0.50 as the number of subjects in that study approached the number of people in the population. Similarly, the studies which employ a 20 mg. dose are all estimating the same parameter, which is the impact of the high dose in this specific population. If the true effect size for this dose is 0.60, then the observed effect size in any of these studies would approach 0.60 as the number of subjects in that study approached the number of people in the population. Since the studies within each subgroup are all estimating the same parameter, the fixed-effect model applies.



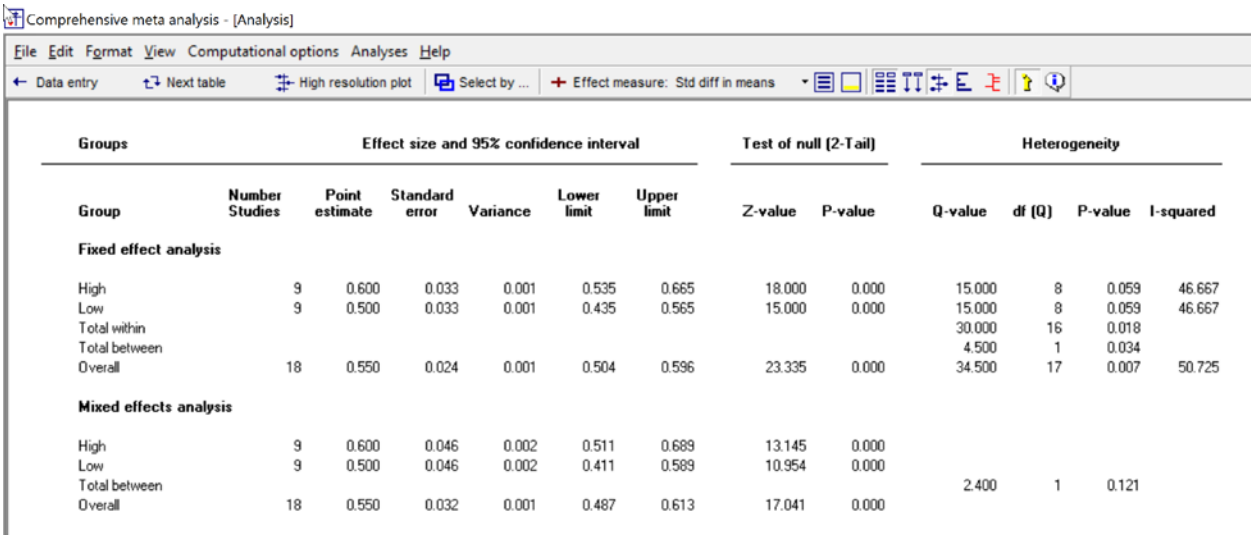
Model	Group by Group	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	High	JJ	0.400	0.100	0.010	0.204	0.596	4.000	0.000
	High	KK	0.450	0.100	0.010	0.254	0.646	4.500	0.000
	High	LL	0.500	0.100	0.010	0.304	0.696	5.000	0.000
	High	MM	0.550	0.100	0.010	0.354	0.746	5.500	0.000
	High	NN	0.600	0.100	0.010	0.404	0.796	6.000	0.000
	High	OO	0.650	0.100	0.010	0.454	0.846	6.500	0.000
	High	PP	0.700	0.100	0.010	0.504	0.896	7.000	0.000
	High	QQ	0.750	0.100	0.010	0.554	0.946	7.500	0.000
	High	RR	0.800	0.100	0.010	0.604	0.996	8.000	0.000
Fixed	High		0.600	0.033	0.001	0.535	0.665	18.000	0.000
	Low	AA	0.300	0.100	0.010	0.104	0.496	3.000	0.003
	Low	BB	0.350	0.100	0.010	0.154	0.546	3.500	0.000
	Low	CC	0.400	0.100	0.010	0.204	0.596	4.000	0.000
	Low	DD	0.450	0.100	0.010	0.254	0.646	4.500	0.000
	Low	EE	0.500	0.100	0.010	0.304	0.696	5.000	0.000
	Low	FF	0.550	0.100	0.010	0.354	0.746	5.500	0.000
	Low	GG	0.600	0.100	0.010	0.404	0.796	6.000	0.000
	Low	HH	0.650	0.100	0.010	0.454	0.846	6.500	0.000
	Low	II	0.700	0.100	0.010	0.504	0.896	7.000	0.000
Fixed	Low		0.500	0.033	0.001	0.435	0.565	15.000	0.000

Figure 45

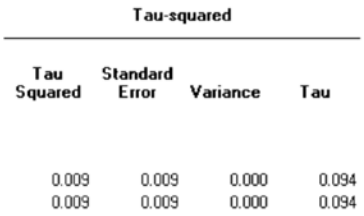


Model	Group by Group	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Fixed	High		0.600	0.033	0.001	0.535	0.665	18.000	0.000
Fixed	Low		0.500	0.033	0.001	0.435	0.565	15.000	0.000

Figure 46



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												
High	9	0.600	0.033	0.001	0.535	0.665	18.000	0.000	15.000	8	0.059	46.667
Low	9	0.500	0.033	0.001	0.435	0.565	15.000	0.000	15.000	8	0.059	46.667
Total within									30.000	16	0.018	
Total between									4.500	1	0.034	
Overall	18	0.550	0.024	0.001	0.504	0.596	23.335	0.000	34.500	17	0.007	50.725
Mixed effects analysis												
High	9	0.600	0.046	0.002	0.511	0.689	13.145	0.000				
Low	9	0.500	0.046	0.002	0.411	0.589	10.954	0.000				
Total between									2.400	1	0.121	
Overall	18	0.550	0.032	0.001	0.487	0.613	17.041	0.000				



	Tau-squared			
	Tau Squared	Standard Error	Variance	Tau
	0.009	0.009	0.000	0.094
	0.009	0.009	0.000	0.094

Figure 47

The effect size and standard error within each subgroup are computed using fixed-effect weights. The effect size for 10 mg. is 0.50, and the effect size for 20 mg is 0.60, so the difference between them is 0.10. To build a confidence interval about this difference, or to test the null hypothesis that the difference is zero, we would use the standard error the difference. This is computed as

$$SE_{DIFF} = \sqrt{V_{M1} + V_{M2}} \tag{1.45}$$

which here is

$$SE_{DIFF} = \sqrt{0.0011 + 0.0011} = \sqrt{0.0022} = 0.0471 \tag{1.46}$$

The 95% confidence interval is for the difference in summary effects is then

$$\begin{aligned} LL_{DIFF} &= M - 1.96 \times SE_{DIFF} \\ UL_{DIFF} &= M + 1.96 \times SE_{DIFF} \end{aligned} \tag{1.47}$$

Which here is

$$\begin{aligned} LL_{DIFF} &= 0.1000 - 1.96 \times 0.0471 = 0.0076 \\ UL_{DIFF} &= 0.1000 + 1.96 \times 0.0471 = 0.1924 \end{aligned} \tag{1.48}$$

Similarly, we could use a *Z*-test to test the null hypothesis that the mean effect size is the same in the two subgroups. Specifically,

$$Z = \frac{M_2 - M_1}{SE_{DIFF}} \tag{1.49}$$

Which here is

$$Z = \frac{0.6000 - 0.5000}{0.0471} = 2.1213 \tag{1.50}$$

With a corresponding *p*-value of 0.034.

The program reports a *Q*-test (which works with any number of subgroups), rather than a Z-test (which only works with two subgroups). When there are two subgroups, the *Q*-value is simply *Z*-squared. Here, *Q* is 4.500, with 1 degree of freedom and a corresponding p-value of 0.034.

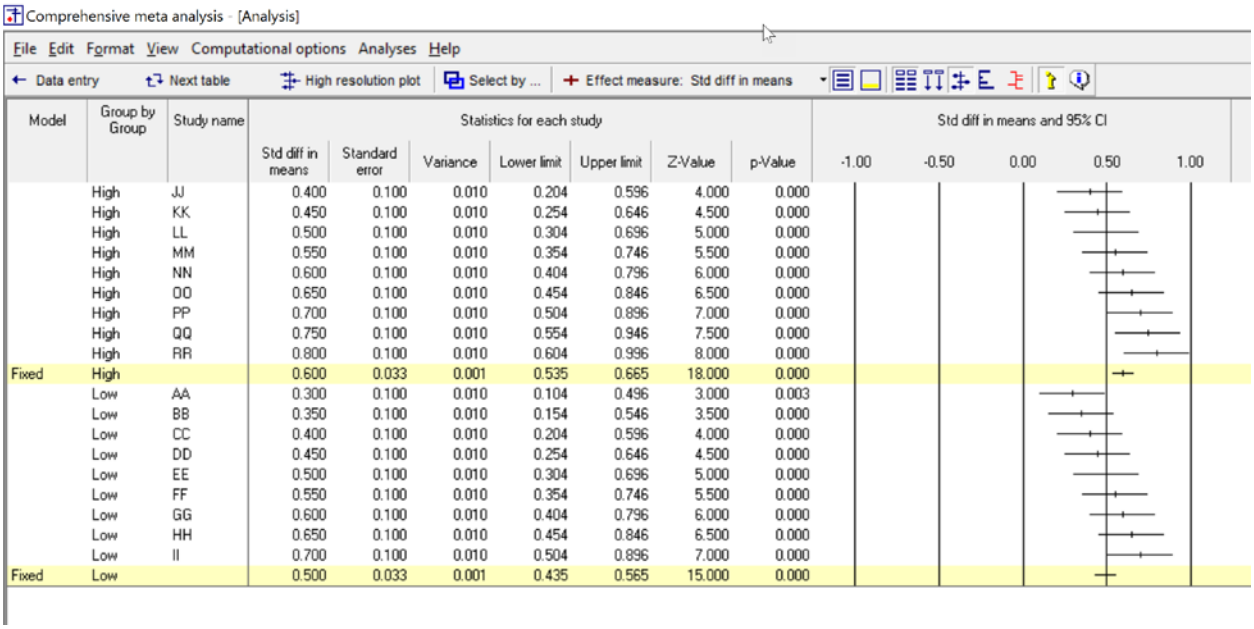
We conclude that the higher dose is more effective than the lower dose. However, this conclusion applies only to the specific population from which the patients were sampled, and the specific protocol employed in these studies.

We say nothing about variation in effect size across studies within subgroups. All studies within a subgroup are estimating the same parameter, and so it makes no sense to ask how this one parameter varies. Rather, all variation in the observed effects within a subgroup must reflect sampling error alone.

Random effects sampling frame

Consider the same example as above, where studies are drawn from different populations, and in each study patients are randomized to either the drug or a placebo. In this case, however, some studies employed a dose of 10 mg. and others employed a dose of 20 mg.

The studies which employ a dose of 10 mg. are not estimating the same parameter. Rather, each is estimating the effect in a specific population, and the effect probably varies from one population to the next. Suppose the mean effect size for this dose is 0.50. If we enrolled the full population in each hospital, the observed effect size for all these studies would not converge on 0.50. Rather, the effects would vary about this value. The same idea applies to the studies which employ a dose of 20 mg. Since the studies within each subgroup are estimating different parameters, the random-effects model applies.



Model	Group by Group	Study name	Statistics for each study: Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	High	JJ	0.400	0.100	0.010	0.204	0.596	4.000	0.000
	High	KK	0.450	0.100	0.010	0.254	0.646	4.500	0.000
	High	LL	0.500	0.100	0.010	0.304	0.696	5.000	0.000
	High	MM	0.550	0.100	0.010	0.354	0.746	5.500	0.000
	High	NN	0.600	0.100	0.010	0.404	0.796	6.000	0.000
	High	OO	0.650	0.100	0.010	0.454	0.846	6.500	0.000
	High	PP	0.700	0.100	0.010	0.504	0.896	7.000	0.000
	High	QQ	0.750	0.100	0.010	0.554	0.946	7.500	0.000
	High	RR	0.800	0.100	0.010	0.604	0.996	8.000	0.000
Fixed	High		0.600	0.033	0.001	0.535	0.665	18.000	0.000
	Low	AA	0.300	0.100	0.010	0.104	0.496	3.000	0.003
	Low	BB	0.350	0.100	0.010	0.154	0.546	3.500	0.000
	Low	CC	0.400	0.100	0.010	0.204	0.596	4.000	0.000
	Low	DD	0.450	0.100	0.010	0.254	0.646	4.500	0.000
	Low	EE	0.500	0.100	0.010	0.304	0.696	5.000	0.000
	Low	FF	0.550	0.100	0.010	0.354	0.746	5.500	0.000
	Low	GG	0.600	0.100	0.010	0.404	0.796	6.000	0.000
	Low	HH	0.650	0.100	0.010	0.454	0.846	6.500	0.000
	Low	II	0.700	0.100	0.010	0.504	0.896	7.000	0.000
Fixed	Low		0.500	0.033	0.001	0.435	0.565	15.000	0.000

Figure 48

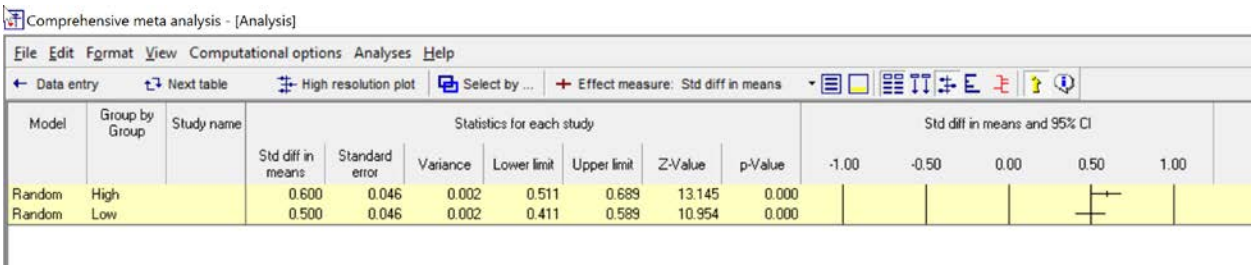


Figure 49

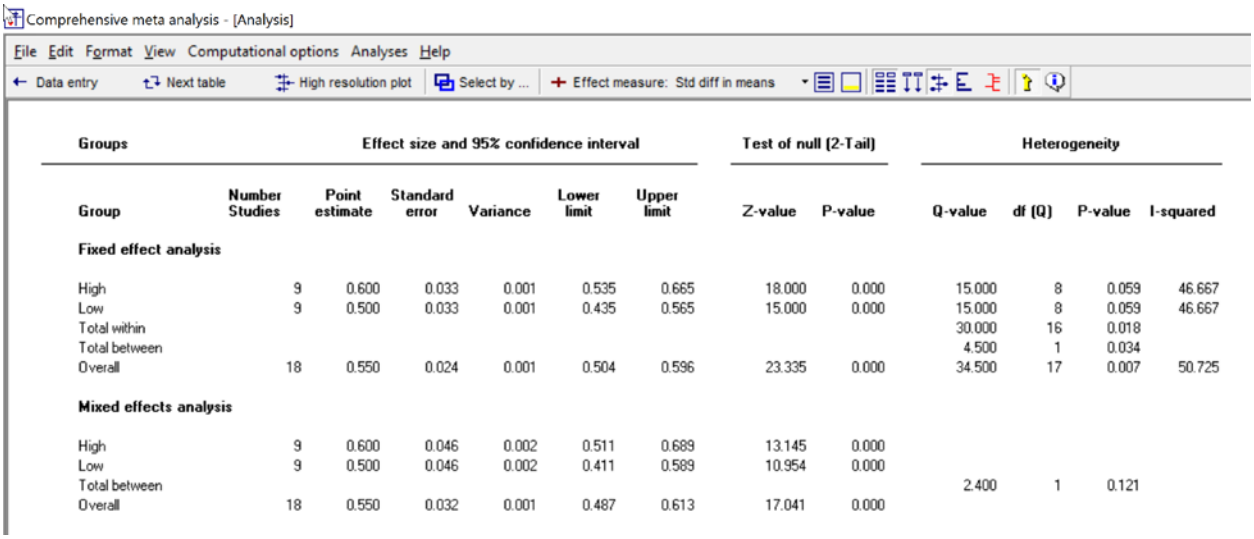


Figure 50

The effect size and standard error within each subgroup are computed using fixed-effect weights. The effect size for 10 mg. is 0.50, and the effect size for 20 mg is 0.60, so the difference between them is 0.10. To build a confidence interval about this difference, or to test the null hypothesis that the difference is zero, we would use the standard error the difference. This is computed as

$$SE_{DIFF} = \sqrt{V_{M1} + V_{M2}} \tag{1.51}$$

which here is

$$SE_{DIFF} = \sqrt{0.0021 + 0.0021} = \sqrt{0.0042} = 0.0646 \tag{1.52}$$

The 95% confidence interval is for the difference in summary effects is then

$$\begin{aligned} LL_{DIFF} &= M - 1.96 \times SE_{DIFF} \\ UL_{DIFF} &= M + 1.96 \times SE_{DIFF} \end{aligned} \tag{1.53}$$

Which here is

$$LL_{DIFF} = 0.1000 - 1.96 \times 0.0646 = -0.0265$$
$$UL_{DIFF} = 0.1000 + 1.96 \times 0.0646 = 0.2265 \quad (1.54)$$

For a test of the null hypothesis that the two doses are equally effective, the Q-value is 2.4000 with 1 degree of freedom and a corresponding p-value of 0.1213

Equivalently (since there are only two subgroups) we could compare the subgroups using a Z-test. Specifically,

$$Z = \frac{M_2 - M_1}{SE_{DIFF}} \quad (1.55)$$

Which is

$$Z = \frac{0.6000 - 0.5000}{0.0646} = 1.5492 \quad (1.56)$$

With a corresponding p-value of 0.1213. Note that (with one degree of freedom) Q is simply Z^2, so 1.5492^2 is equal to 2.4000.

We conclude that there is no evidence that the higher dose is more effective than the lower dose. If there was evidence that the higher dose was more effective, the conclusion would apply to the two universes from which these two sets of studies were sampled.

We would also estimate the variation in effect size across studies within subgroups. The standard deviation of true effects within subgroups is estimated as 0.0935. In the universe of populations from which these studies were sampled, most studies that employed 10 mg, will have a true effect size in the range of

$$PI_{LL} = d - 2T = 0.50 - 2 \times 0.0935 = 0.313$$
$$PI_{UL} = d + 2T = 0.50 + 2 \times 0.0935 = 0.687 \quad (1.57)$$

and most that employed 20 mg will have a true effect size in the range of

$$PI_{LL} = d - 2T = 0.60 - 2 \times 0.0935 = 0.413$$
$$PI_{UL} = d + 2T = 0.60 + 2 \times 0.0935 = 0.787 \quad (1.58)$$

Regression

Consider a regression analysis, where the dose of drug varied by study. Our goal is to assess the relationship between dose and effect size.

- The fixed-effect model applies if all studies at any given dose are drawn from a single population. The studies at a given dose share a common *true* effect-size, and so the *true* effect size is *fixed*, or constant.

- The random-effects model applies if the studies at a given dose are drawn from a universe of populations. The *true* effect size for any given dose varies from one population to the next, and the studies are sampled at *random* from this universe.

Fixed-effect sampling frame

Consider the same example as above, where all studies are drawn from the same population, and in each study patients are randomized to either the drug or a placebo. In this case, however, the dose of the drug is set (by random assignment) to some value between 10 and 50 mg.

The studies which employ a dose of 10 mg. are all estimating the same parameter, which is the impact of this dose in this specific population. If the true effect size for this dose is 0.50, then the observed effect size in any of these studies would approach 0.50 as the number of subjects in that study approached the number of people in the population. The same idea applies to all other doses. Since the studies with the same dose are all estimating the same parameter, the fixed-effect model applies.

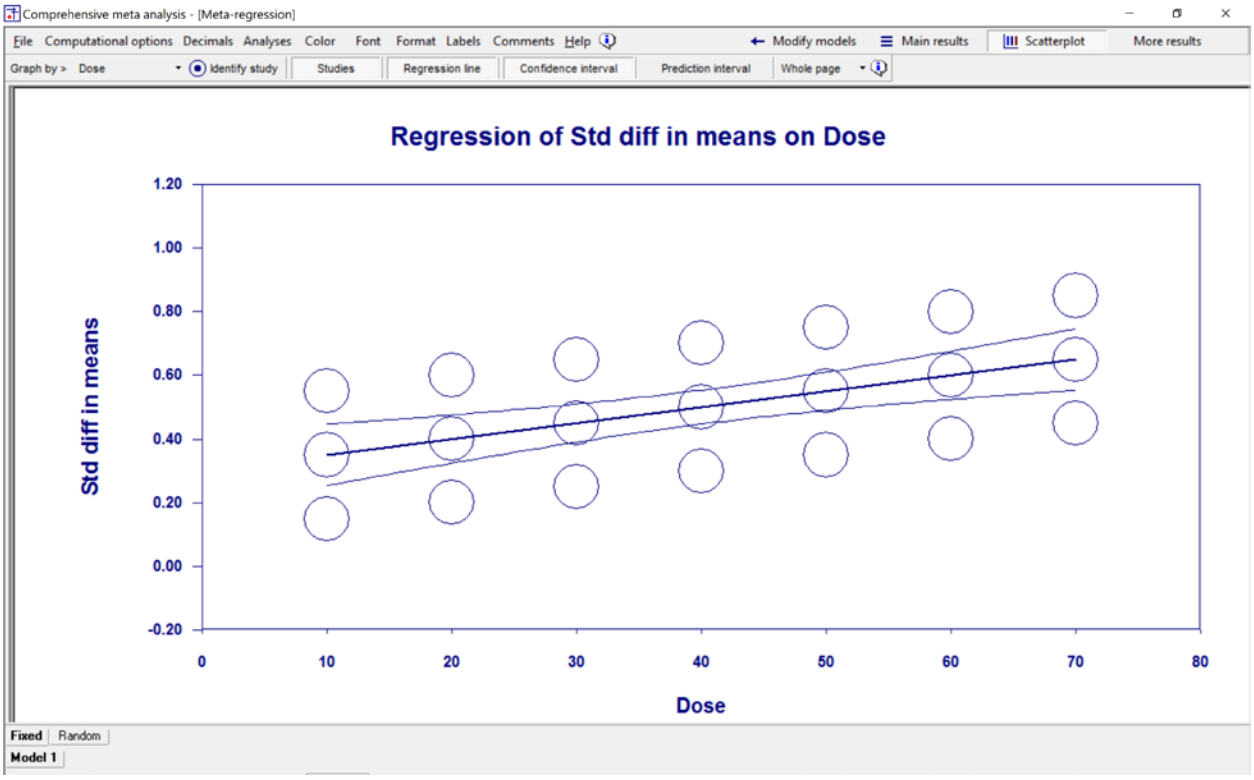


Figure 51

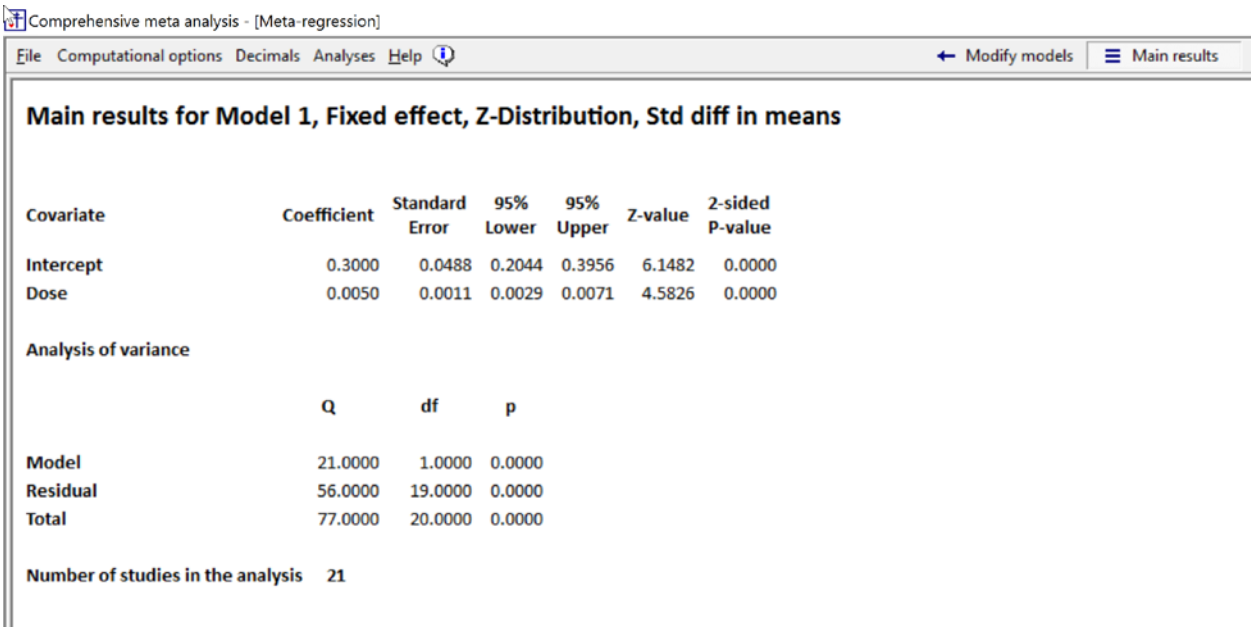


Figure 52

The results of this regression are displayed in Figure 51 and Figure 52. Consider the three studies that employed a dose of 40 mg. We assume that the true effect size for these three studies is the same, and the difference in observed effects is due entirely to sampling error. If we increased the sample size in each of these studies, then as the sample size approached the size of the full population, the observed effect size in the three studies must converge on the same value, since all three studies are estimating the effect size in the same population. By analogy, we can think of each set of three studies as a subgroup. If we displayed these five subgroups (20 mg, 30mg, and so on) using the subgroups framework (**Error! Reference source not found.**), the variation in effects within each subgroup would be due entirely to sampling error.

In this fictional example, we've created three studies at 10 mg, three at 20 mg., and so on. In an analysis with real data, that there may be only one study at any given dose. Nevertheless, the sampling frame tells us that *if there were* multiple studies at the same dose, they would all be estimating the same parameter because they are drawn from the same population.

The coefficient for Dose (B_1) is 0.0050 with a standard error of 0.0011. The 95% confidence interval is given by

$$\begin{aligned} LL_{B1} &= B_1 - 1.96 \times SE_{B1} \\ UL_{B1} &= B_1 + 1.96 \times SE_{B1} \end{aligned} \tag{1.59}$$

Which here is

$$\begin{aligned} LL_{B1} &= 0.0050 - 1.96 \times 0.0011 = 0.0029 \\ UL_{B1} &= 0.0050 + 1.96 \times 0.0011 = 0.0071 \end{aligned} \tag{1.60}$$

If we multiply the coefficient and the effect size by 10, we get the impact of a 10-unit increase in dose. Our best estimate is that for every 10 mg. increase in dose, the effect size increases by 0.05. The confidence interval tells us that the increase could be as low as 0.029 or as high as 0.071. Since this range does not include zero, we can reject the null hypothesis that Dose is unrelated to effect size.

Similarly, for a test of the null hypothesis that Dose is unrelated to effect size, the Z-value is given by

$$Z = \frac{B_1}{SE_{B1}} \tag{1.61}$$

Which here is

$$Z = \frac{0.0050}{0.0011} = 4.5826 \tag{1.62}$$

and the corresponding p-value is < 0.001. Again, we conclude that the higher dose is more effective than the lower dose.

These conclusions – the a higher dose is related to a higher effect size, and that a 10 unit increase in dose will yield an increase of 0.029 to 0.071 – applies only to the specific population from which the patients were sampled, and the specific protocol employed in these studies.

We say nothing about variation in effect size across studies with the same dose. All studies with the same dose are estimating the same parameter, and so it makes no sense to ask how this one parameter varies. Rather, for any given dose, all variation in the observed effects must reflect sampling error alone. For example, based on the regression equation or the plot we would predict that a study with a dose of 40 mg will have an effect size of

$$d = 0.30 + 0.005 \times 40 = 0.50 \tag{1.63}$$

This is only an estimate – since the prediction is imperfect, the actual effect size for a dose of 40 mg is almost certainly different than the predicted value. However, whatever the correct value is, it is the same for all studies at this dose because (by definition) they are all estimating the same value. In this fictional analysis, we have included three studies at this dose. If we were to increase the sample size in each study until it approached the size of the population, the observed effects in these three studies would converge on the same value.

The same idea is easily extended to a regression with multiple covariates. The fixed-effect model applies as long as the studies are identical on all dimensions except those captured by the covariates.

Random-effects sampling frame

Consider the same example as immediately above, where studies are drawn from different populations, and in each study patients are randomized to either the drug or a placebo. In this case, however, the dose in each study was either 10, 20, 30, 40, or 50 mg.

The studies which employ a dose of 10 mg. are not estimating the same parameter. Rather, each is estimating the effect in a specific population, and the effect probably varies from one population to the next. Suppose the mean effect size for this dose is 0.50. If we enrolled the full population in each hospital, the observed effect size for all these studies would *not* converge on 0.50. Rather, the true effects would vary about this value. The same idea applies to studies that used a dose of 20 mg, and so on. Since the studies at any given dose are estimating different parameters, the random-effects model applies.

In a real application, there may be only one study with any given dose. The key point is that *if there were* multiple studies with the same dose, they would each be estimating a unique parameter, because each is drawn from a different population.

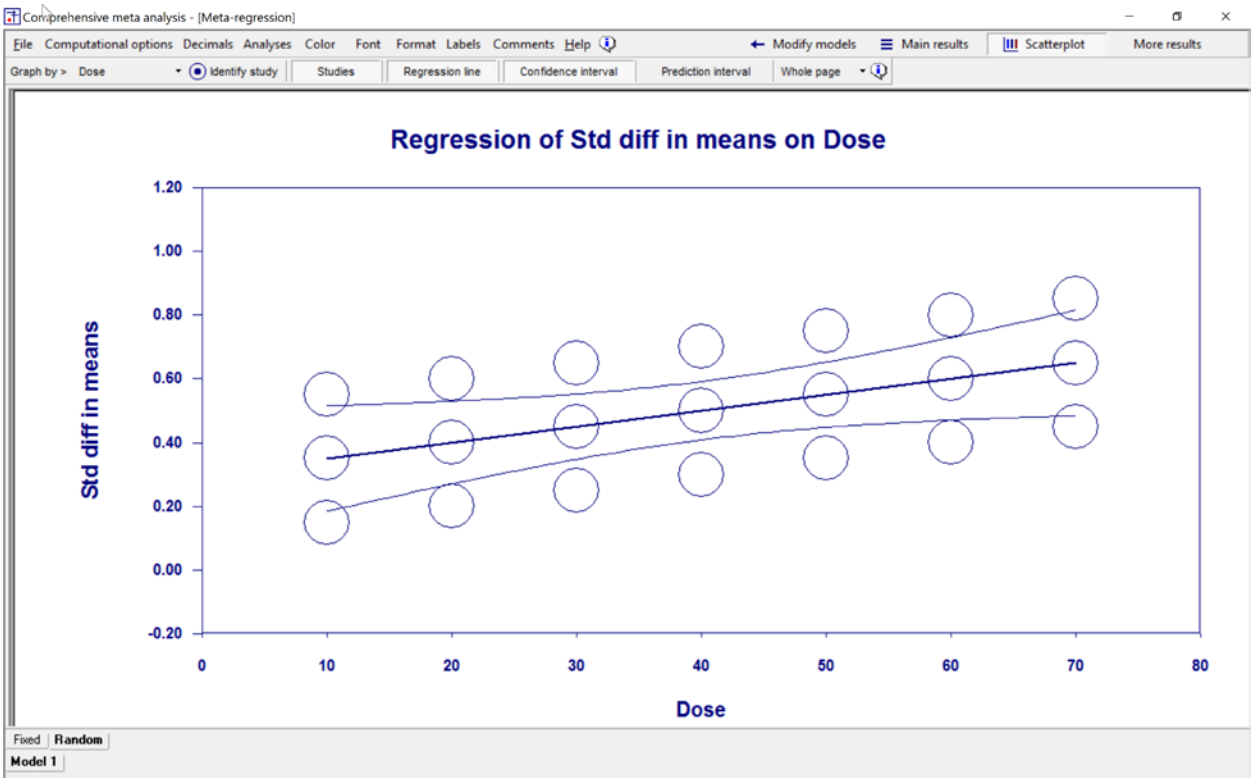


Figure 53

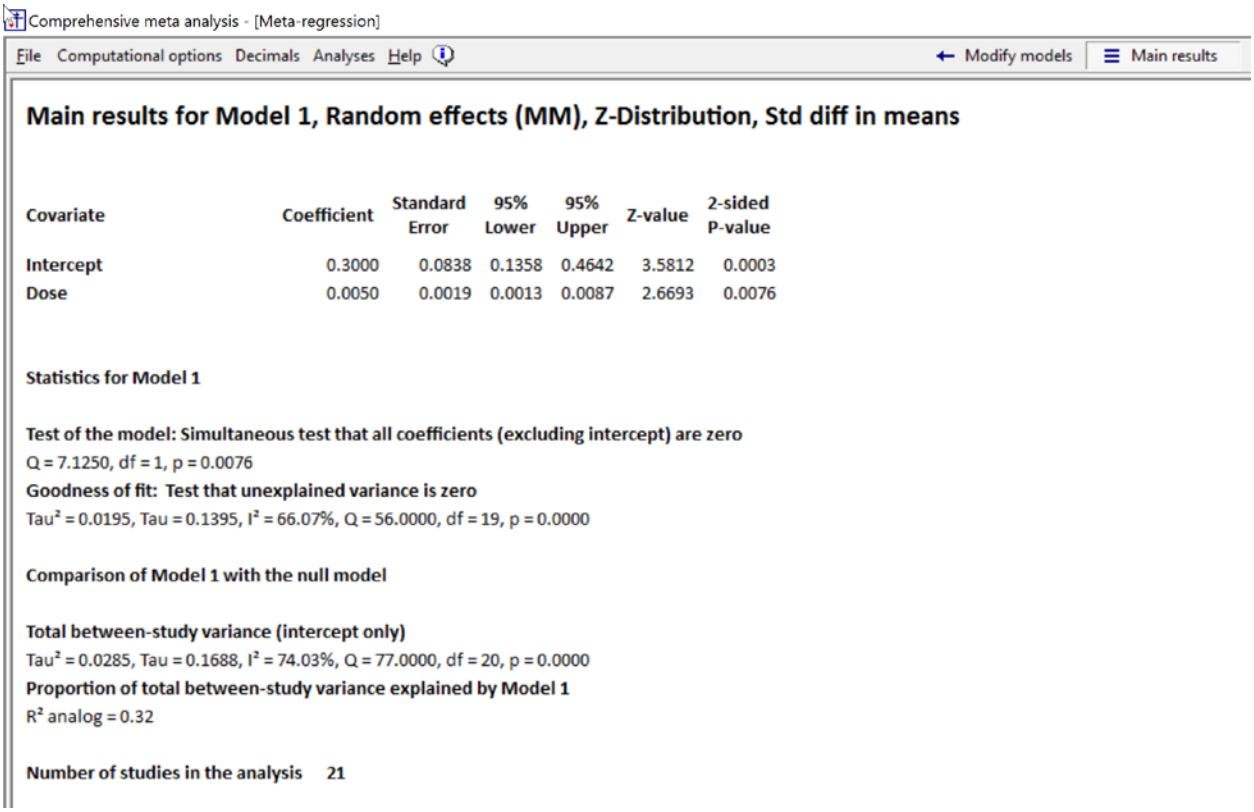


Figure 54

The results of this regression are displayed in Figure 53 and Figure 54. Consider the three studies that employed a dose of 40 mg. We assume that the true effect size for each of these three studies is *not* the same, and the difference in observed effects is due partly to differences in the true effect size and partly to sampling error. If we increased the sample size in each of these studies, then as the sample size approached the size of the full population, the observed effect size in the three studies would *not* converge on the same value, since each of these three studies is estimating the effect size in a unique population.

By analogy, we can think of each set of three studies as a subgroup. If we displayed these five subgroups (20 mg, 30 mg, and so on) using the subgroups framework (Figure 48), the variation in effects within each subgroup would be due partly to differences in the true effects and partly to sampling error.

The coefficient for Dose (B_1) is 0.0050 with a standard error of 0.0019. The 95% confidence interval is given by

$$\begin{aligned} LL_{B1} &= B_1 - 1.96 \times SE_{B1} \\ UL_{B1} &= B_1 + 1.96 \times SE_{B1} \end{aligned} \tag{1.64}$$

which here is

$$LL_{B1} = 0.0050 - 1.96 \times 0.0019 = 0.0013$$
$$UL_{B1} = 0.0050 + 1.96 \times 0.0019 = 0.0087 \quad (1.65)$$

If we multiply the coefficient by 10 we get the impact of a 10-unit increase in dose. Our best estimate is that for every 10 unit increase in dose, the effect size increases by 0.05. The actual increase could be as low as 0.013 or as high as 0.087. Since this range does not include zero, we can reject the null hypothesis that Dose is unrelated to effect size.

Similarly, for a test of the null hypothesis that Dose is unrelated to effect size, the Z-value is given by

$$Z = \frac{B_1}{SE_{B1}} \quad (1.66)$$

Which here is

$$Z = \frac{0.0050}{0.0019} = 2.6693 \quad (1.67)$$

and the corresponding p-value is = 0.0076. Again, we conclude that the higher dose is more effective than the lower dose.

Since there is only one covariate in the model, the test of the model is identical to the test of the covariate. The only difference is that the test of the model employs *Q* rather than *Z*, where *Q* is simply *Z* squared. Here, *Q* is 7.1250 with 1 degree of freedom and *p* = 0.0076.

These conclusions – that a higher dose is related to a higher effect size, and that a 10 unit increase in dose will yield an increase of 0.013 to 0.087 – applies not only to the specific populations included in the analysis, but to the universe of populations from which these studies were sampled.

We are also interested in the variation in effect size across studies with the same dose. If there are multiple studies at a given dose, we want to know if the effect size for all of them falls within a narrow interval or varies substantially. This is addressed by the standard deviation of true effects (*T*), which is reported as 0.1395. If the mean true effect size for a given dose is 0.60, then at this dose, most populations would have a true effect size in the range of

$$PI_{LL} = d - 2T = 0.60 - 2 \times 0.0195 = 0.561$$
$$PI_{UL} = d + 2T = 0.60 + 2 \times 0.0195 = 0.639 \quad (1.68)$$

This formula assumes that the mean effect size and *T* are known. The program will compute a prediction interval that takes into account the uncertainty in these estimates.

The same idea is easily extended to a regression with multiple covariates. The fixed-effect model applies if the studies in the analysis are identical on all dimensions except those captured by the covariates. The random-effects model applies otherwise, as it does in this example.

PUTTING IT IN CONTEXT

To this point we've presented a simple analysis, a subgroups analysis, and a meta-regression using both a fixed-effect model and a random-effects model.

Actually, all of these cases can be subsumed under a common rule. If all variation about the predicted effect size is due to sampling error, then the fixed-effect model applies. Otherwise the random-effects model applies. In the simple case, the predicted effect for all studies in the grand mean. In the case of subgroups, the predicted effect for each study is the mean of that study's subgroup. In regression, the predicted effect for each study is the corresponding point on the regression line. We present a series of fictional examples to illustrate these points.

The simple analysis was presented in Figure 42 for the fixed-effect model, and in Figure 43 for the random-effects model. The two analyses are the same, except that the standard error of the summary effect size is larger under the random-effects model. This affects the confidence interval for the summary effect, and the test of the null hypothesis that the summary effect size is zero. These are labeled [A] and [B] in each figure.

The subgroups analysis was presented in Figure 46 using the fixed-effect model within subgroups, and Figure 49 using the random-effects model within subgroups. The standard error of the difference between subgroups is larger under the random-effects model than under the fixed-effect model. This affects the confidence interval for the difference between subgroup, and the test of the null hypothesis that the difference between subgroups is zero (Figure 47 and Figure 50).

The regression analysis was presented in Figure 52 using the fixed-effect model within subgroups, and Figure 54 using the random-effects model within subgroups. The standard error of the coefficient for Dose is larger under the random-effects model than under the fixed-effect model. This affects the confidence interval for the coefficient, and the test of the null hypothesis that the coefficient is zero.

The reason that the standard error is consistently larger under the random-effects model is most easily explained for the simple analysis, as follows. In our example, the standard error of the summary effect under the fixed-effect model is

$$SE_d = \sqrt{\frac{V}{N}} \ , \tag{1.69}$$

And under the random-effects model is

$$SE_d = \sqrt{\frac{V}{N} + \frac{T^2}{k}} \ , \tag{1.70}$$

These are conceptual formulae, since in any real analysis the V will vary, but they highlight the difference between the two models.

Under the fixed-effect model we are estimating the effect size in one population. The error term reflects the fact that we are generalizing from the samples in these studies to the full population.

Under the random-effects model we are estimating the *mean* effect size in a universe of populations. The first part of the error term reflects the fact that we are generalizing from the subjects in each study to the full population in that study. The second part of the error term reflects the fact that we are generalizing from the populations in our analysis to the universe of populations from which these studies were sampled.

It's this second component that allows us to generalize beyond the studies in the analysis, to a wider universe of studies. The penalty is that we have a larger error term, but the advantage is that we get to generalize to a wider universe.

In our examples the estimate of the effect size itself (in the simple analysis), of the difference between subgroups (in the subgroup analysis) and of the coefficient for Dose (in the regression) was the same for both statistical models. This was true in our fictional examples because we assigned the same variance to all studies. In any real analysis, these values will be different under the two models.

COMMON MISTAKES

Comparing the results under the two models

Researchers sometimes ask why the results differ under the two models. This question tends to take on some urgency when the random-effects model yields a non-significant *p*-value for some test while the fixed-effect model yields a *p*-value well under 0.05.

The most relevant answer is that we should not be comparing the two models. The fixed-effect model asks "Based on these samples from one population, what can we say about the common effect in this population?" The random-effects model asks "Based on these studies from a sample of populations, what can we say about the mean effect size in the universe of similar populations?" The two models are asking two fundamentally different questions, and in general there would be no reason to compare them.

We did compare them here, because one of our goals is to show how the logic of each model leads to a set of rules for assigning weight to each study, and how these weights affect the statistics. This is appropriate in the context of this volume, but not in the context of a real analysis.

How to choose a statistical model

There is a widespread belief that we should start any analysis using the fixed-effect model, and switch to the random-effects model if there is clear evidence of variance in the effects. This approach is incorrect. The choice of a model should be based on our knowledge of the sampling frame, and not on a test for heterogeneity.

If the studies have been pulled from the literature, then (in almost all cases) logic tells us that the effect size probably varies from study to study, and we should use the random-effects model. This is simply the way things are. If we are looking at the impact of an intervention, that impact will almost always vary from one population to the next (or as details of the protocol change from one study to the next). The variation might be trivial, but once there is any variation the random-effects model applies.

When T^2 is estimated as zero

A related issue is the case where we are using the random-effects model, and it turns out that T^2 is estimated as zero. Since the study weight under the fixed-effect model is

$$W_i = \frac{1}{V_i} \tag{1.71}$$

and the study weight under the random-effects model is

$$W_i = \frac{1}{V_i + T^2}, \tag{1.72}$$

when T^2 is estimated as zero, the two models use the same weights and yield precisely the same results. In this event, researchers sometimes report that they have reverted to the fixed-effect model.

Again, this is a mistake. If we are using the random-effects model and T^2 is estimated as zero, then we are still using the random-effects model – it just so happens that another (different) model would yield the same results. In fact, in a case where the random-effects model is called for, and T^2 is estimated as zero, it's virtually certain that the actual value of T^2 is greater than zero, and our estimate is simply too low. Indeed, when T^2 is estimated as zero, the *computed* value is less than zero which we know is too low (since it's not possible to have a negative variance). We simply reset it to zero in order to proceed with the computations.

Problems with the random-ef fects model

While the random-effects model is typically the correct model to use, there are often problems implementing this model in a real-world analysis.

In order to implement this model with reasonable accuracy, we need a sufficiently precise estimate of T^2, and this requires a certain number of studies. If we don't have enough studies, then the results will be suspect. The extent of this problem is likely to vary by field of research. If we are working in a field where the heterogeneity is low, then we may be able to get a reasonably precise estimate of T2 even with a small number of studies. However, if we are working in a field where the effect size varies substantially across studies, attempts to estimate T2 with a small number of studies are likely to be problematic.

The random-effects model assumes that the effects in our sample of studies are a random sample of the effects in the relevant universe. In order to properly meet this assumption, we would need to be clear about how we define the relevant universe. And, we would need ensure that our studies are actually a random sample from this universe. We are likely to fall short on both counts, and we need to consider how much this affects the validity of our results.

Causal vs. observational relationships

In the example that we used in this chapter for the fixed-effect model, the finding that the effect size was related to dose in the subgroups comparison and the regression would be interpreted as a causal, rather than an observational relationship. That's because we started with one population and randomly assigned studies to a dose of the drug. However, if the studies at one dose were based on one population while those at another dose were based on a different population, a finding that effect size was related to dose would be observational rather than causal, since it could be due to factors that are confounded with dose. This holds true even when we are using the fixed-effect model.

In the example that we used in this chapter for the random-effects model, a finding that the effect size was related to dose in the subgroups comparison and the regression would be interpreted as an observational relationship, since it could be due to factors that are confounded with dose. This would hold true in almost any random-effects analysis. Since the population varies from study to study, there

will almost always be a concern that dose is confounded with other factors. This is discussed more fully in another chapter.

CAVEATS

The examples we used to illustrate the fixed-effect model were designed to parallel those we used to illustrate the random-effects model. In practice, if a drug company wanted to design RCTs to compare two (or more) doses of a drug, it would be better to employ a design where each study included all doses and placebo, rather than using one dose in each study. This would allow for a head-to-head comparison of the doses. This a more powerful approach than comparing each dose to a placebo. Additionally, if the analysis shows that one drug is more powerful than the other in a head-to-head comparison, this would be causal rather than observational.

DISPLAYING THE RESULTS

Figure ____ shows the subgroups comparison using the fixed-effect model [A] and using the random-effects model [B]. We show how to partition the sum of squares for [A] but not for [B]. The reason is as follows.

Under the fixed-effect model, the weight (W_i) assigned to each study is

$$W_i = \frac{1}{V_i} \tag{1.73}$$

Since the variance for any study (V_i) is the same in all analyses, the weight assigned to any given study, and the total sum of squares, is constant. This is why

$$Q_{Total} = Q_{Between} + Q_{Error} \tag{1.74}$$

And we can partition the sum of squares.

By contrast, under the random-effects model, the weight assigned to each study is

$$W_i = \frac{1}{V_i + T^2} \tag{1.75}$$

Where T2 is our estimate of the between-study variance. The value of T2 changes depending on the context. Typically, T2 across all studies is larger than T2 within subgroups. Therefore,

$$Q_{Total} \neq Q_{Between} + Q_{Error} \tag{1.76}$$

And we cannot partition the sum of squares. Therefore, for the random-effects statistics [B], we display $Q_{Between}$ only. This is the statistic that addresses the difference between subgroup means.

The same logic applies in the case of a regression. When we use the fixed-effect model we can partition the variance into its component parts as shown in _____, because the weight assigned to any given study is constant, and therefore ____ applies. By contrast, when we use the random-effects model, the weight assigned to any given study varies changes as we move from the initial analysis to the regression. In this case ____ applies and we cannot partition the sum of squares.

NEED THIS

R2

In the case of the RE model, the fact that the observed effects do not fall in the regression line is due to three items – the mean effects does not fall directly on the line, the true effect in each study does not fall at the mean, and the observed effect for each study is not the same as the true effect for that study.

Denominator is variance of true effects about the mean

Numerator is ___

To compute R2 we remove the last item

In the case of the FE model, the fact that the observed effects do not fall on the regression line is due to two items – the common effect does not fall directly on the line, and the observed effect for each study is not the same as the true effect for that study.

Denominator is _____

Numerator is _____

A THIRD STATISTICAL MODEL

In this chapter we've focused on two statistical models – the fixed-effect model and the random-effects model. There is a third possible model as well. This model is sometimes called the fixed-effects model, where the word "effects" is in the plural rather than the singular.

This model would apply in the case where the studies in the analysis are from different populations, so (as in the random-effects model) we are trying to estimate a mean effect rather than a common effect. However, (unlike the random-effects model) our goal is to estimate the mean in this specific set of studies only, and not to extrapolate beyond them. The word "fixed" means "fixed" in the sense of "set" or "defined" – we are interested in this fixed set of studies only. In fact, this use of the term (to mean "defined" rather than "common" is the usual use of the term in many fields.

As it turns out, the weights we would assign under this model are identical to the weights we would assign under the fixed-effect model. And, the standard error of the mean under this model would be the same as the standard error of the common effect size under the fixed-effect model. Recall that the standard error under the fixed-effect model (assuming all studies share a common variance) is

$$SE_M = \sqrt{\frac{V}{N}} \tag{1.77}$$

Whereas under random-effects, is it

$$SE_M = \sqrt{\frac{V}{N} + \frac{T^2}{k}} \tag{1.78}$$

The second term under the radical reflects the fact that we are using the mean in our sample of studies to generalize to a wider universe of studies. In the fixed-effect model (singular) this term is zero because we are not generalizing to a wider universe – we are limiting ourselves to one population. In the fixed-effects (plural) model this term is zero because we are not generalizing to a wider universe – we are simply describing the mean for the studies in our analysis.

This model is only rarely appropriate for a meta-analysis. One reason is that when we publish an analysis, readers will generalize the results to other populations. If the confidence interval and test of significance are based on (1.77) and the results are generalized in this way, then they are meaningless.

In fact, the same idea applies when we are using a fixed-effect model – the standard error assumes that we are not generalizing beyond this one population. When the fixed-effect model actually applies, this is a plausible restriction and one that we can make explicitly. Under the fixed-effects model it's less plausible and less likely to be followed.

A second reason that the fixed-effects model (plural) is rarely appropriate is that under this model larger studies may dominate the analysis. This makes sense under the fixed-effect model, since all studies are estimating the same parameter. If one study has ten times as much information about that parameter than another, it should get ten times as much weight. By contrast, under the fixed-effects model, this may not be appropriate. If we are working with a number of schools, and the larger schools enrolled

more students, then this approach might make sense. In this case the mean effect in the analysis might reflect the mean for all students in a district. But in most cases (especially when studies are being pulled from the literature), this is not likely to be the case.

For these reasons, we address this model only as an afterthought. In the event that someone wanted to use this model, then they could select the fixed-effect option in any software and the results would apply.

The statistical models typically employed in meta-analysis are the fixed-effect model and the random-effects model.

If all studies are based on the same population and are identical in all respects, the fixed-effect model applies. In this case our goal is to estimate the common effect size in this population. By contrast, if studies are based on different populations, the random-effects model applies. In this case our goal is to estimate the mean (rather than the common) effect size, and also to estimate how the effect size varies across populations.

While the distinction between the two models is typically discussed in the context of a simple analysis, it also applies more generally. In a subgroup analysis, the fixed-effect model applies only if all studies within a subgroup are based on the same population. In a meta-regression, the fixed-effect model applies only if all studies with the same value on all factors are drawn from the same population. These conditions are rarely met, and therefore the random-effects model is almost always a better fit for the data.

There is a common belief that the selection of a model should be based on a test for heterogeneity, but this is incorrect. The selection of a model must be based on the sampling frame. If we apply the fixed-effect model when the random-effects model is called for, the results could lead to serious errors in inference.

FIXED-EFFECT ANALYSIS

Here, we show a regression analysis to assess the relationship between Dose and effect size for the ADHD studies. In the ADHD example, the studies were pulled from the literature, and the population varies from study to study. Therefore, it's clear that the fixed-effect model does not apply in this case. What follows is an academic exercise to show how we would proceed if the fixed-effect model did apply – that is, if all studies had sampled patients from the same population, and were identical to each other on all factors except for the one factor under being studied in this analysis. We will show three examples.

SUD
DOSE
SUD + DOSE
START WITH SUD – SHOW AS SUBGROUPS
THEN ADD DOSE
FOLLOW SAME MODEL AS FOR RE

To navigate to this screen Click [Run regression] [A]

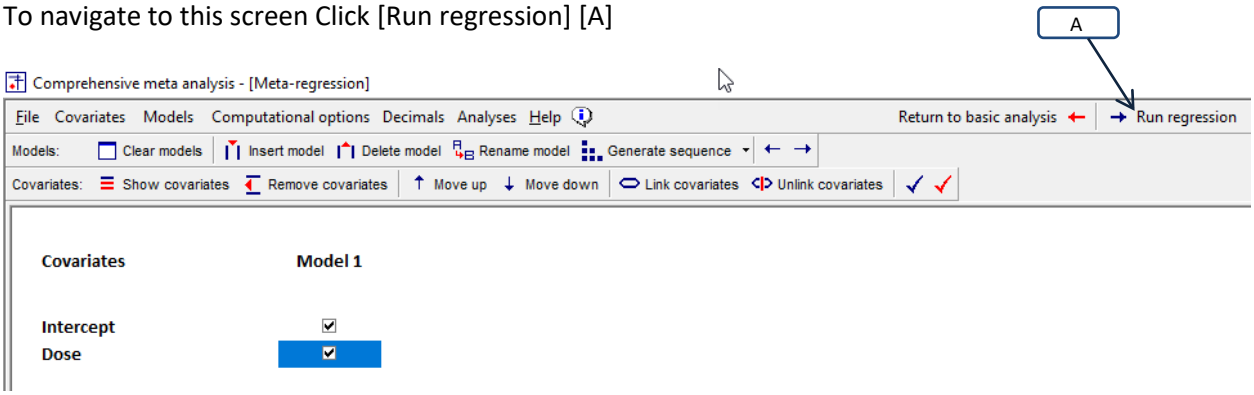


Figure 55 | Setup

The toolbar changes as shown in Figure 56.

- Click "Main results" [B]
- Click [Fixed] [C]

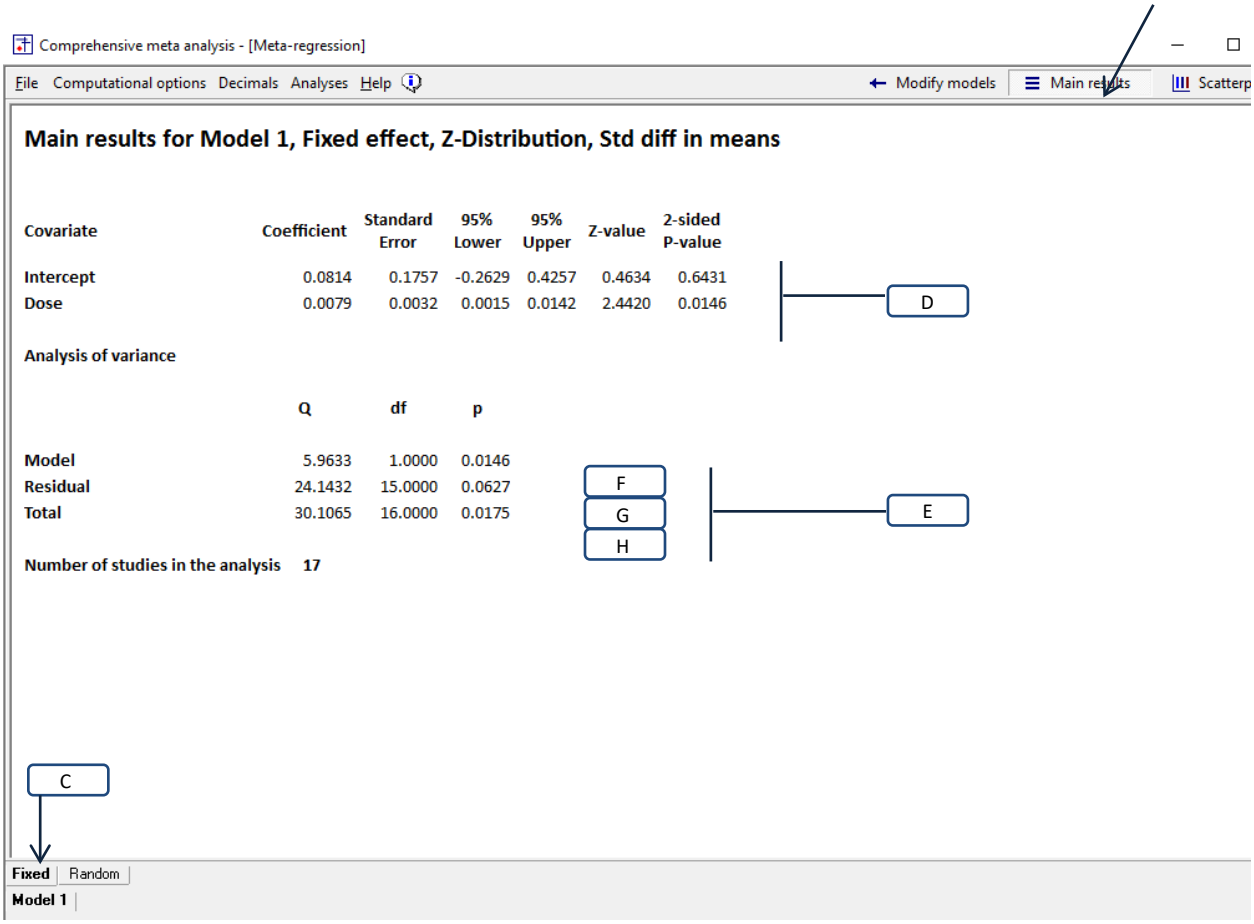


Figure 56 | Main results | Fixed-effect

WHAT DOES IT MEAN – WHAT ARE WE LOOKING AT IN THE RESULTS

WHY WE CAN DECOMPOSE THE VARIANCE HERE

FIXED EFFECTS (PLURAL)

SHOW WHAT HAPPENS TO THE SE IF WE OMIT T2

Can we still talk about variance being explained, as Qbet/QTot

Test of the model

Analysis of variance

In a primary study, the total sum of squares (SS) is the sum of the SS explained by the model and the SS residual. Similarly, in a meta-analysis (with fixed-effect weights) the total weighted sum of squares (Q_{TOTAL}) is the sum of the Q explained by the model and the Q residual. As shown in Figure 56 [E],

$$Q_{TOTAL} = Q_{MODEL} + Q_{RES} = 30.1065 = 5.9633 + 24.1432 \quad (1.79)$$

Similarly, the total *df* is the sum of the model and the residual,

$$df_{TOTAL} = df_{MODEL} + df_{RES} = 16 = 1 + 15 \quad (1.80)$$

Total

The total [H] is the test that the variance for the full set of studies (with no predictors) is zero. The Q-value is 30.1065 with df = 16 and p = 0.0175. This tells us that the true effects vary about the grand mean of all effects.

Model

The model [F] is the test that the predictive model explains *any* of the variance in effect size. Put another way, it asks if the dispersion of effects about the regression line is smaller when the regression line is based on the covariates rather than based solely on the grand mean. Here, Q = 5.9633 with df = 1 and p = 0.0146, so we conclude that the predictive model explains (at least) some of the variance in effect size.

Residual

The residual [G] is the test of the null hypothesis that the *true* effect size for each study lies on the regression line (and that the variation of *observed* effects from the regression line is due to sampling error. The Q value is 24.1432 with df = 15 and p=.0627. Since the criterion alpha for this test is typically 0.10, we would reject the null and conclude that for any point on the regression line, the variance of true effects is larger than would we would expect based on sampling error alone.

Impact of individual covariates [C]

In Figure 56, the test of the model [F] is an omnibus test for the full set of covariates. It tells us that the set as a whole is related to effect size. By contrast, the rows at the top [D] address the *unique* impact of each covariate – that is, the impact of each covariate when *all of the other covariates are* held constant. In this example, where there is only one covariate, the two tests are the same.

Dose

The coefficient for Dose is 0.0079, which means that for every increase of one unit in Dose the predicted effect size (*d*) will increase by 0.0079. Equivalently (but more intuitively) we could multiply both the unit and coefficient by 50, and say that for a 50-unit increase in dose the predicted effect size will increase by some 0.40 standard deviations.

The coefficient (0.0079) plus/minus 1.96 times the standard error (0.0032) yields the 95% confidence interval for the coefficient, which is 0.0015 to 0.0142. The coefficient divided by its standard error yields a *Z* value of 2.4420, and a corresponding *p*-value of 0.0146.

The model

The total *Q* of each effect size about the grand mean can be partitioned into its component parts – the *Q* due to the variation in effect size that can be explained by the covariates, and the part that cannot.

- Model. The *Q*-value for the model is 5.9633 with *df* = 1 and *p* = 0.0146, which tells us that effect size is related to at least one of the covariates (here, to the one covariate).
- Residual. The *Q*-value for the residual is 24.1432 with *df* = 15 and *p*=0.0627, which tells us that the model probably does not explain all the variance in true effects.
- Total. The *Q*-value for the total is 30.1065 with *df* = 16 and *p* = 0.0175, which tells us that that the true effect sizes vary when we ignore covariates and work with deviations of all studies from the grand mean.

Individual covariates

In this example the only covariate is Dose. If there was more than one covariate, the table at the top would addresses the impact of each covariate *with all other covariates held constant*.

COMPUTATIONAL OPTIONS

The program allows you to set various computational options

On the regression screen click Computational Options on the menu.

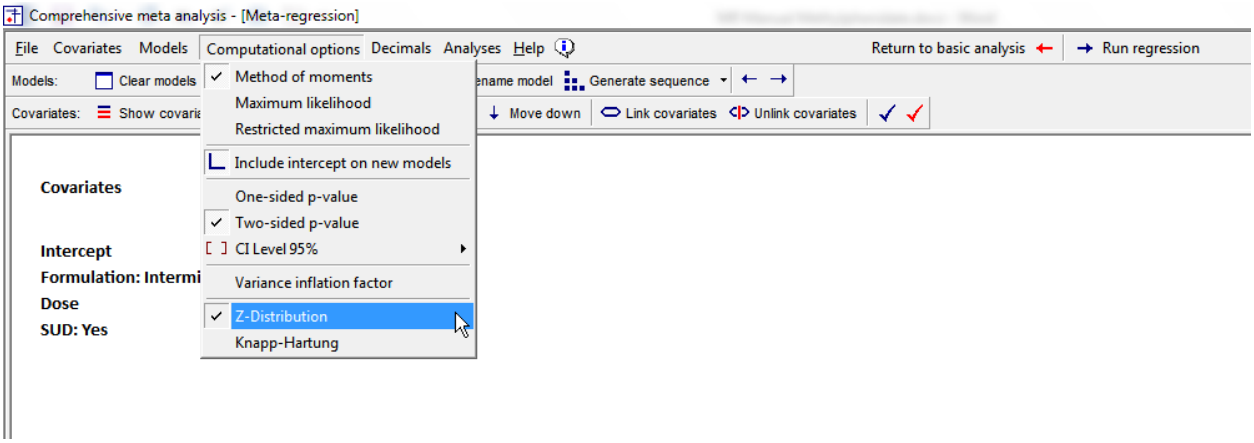


Figure 57 | Regression | Set statistical options

OPTIONS FOR ESTIMATING T^2 (MM, ML, REML)

When we select random-effects the program needs to estimate the value of tau-squared (τ^2), the true between-studies variance. (We use the Greek symbol τ^2 to represent the true value, and T^2 to represent the sample estimate of that value).

There are three approaches commonly used to partition the variance and estimate τ^2. These are

- Method of moments (MM), also known as the DerSimonian and Laird method)
- Unrestricted maximum likelihood (ML) also known as maximum likelihood
- Restricted maximum likelihood (REML)

Each of these methods has advantages and disadvantages.

If we are not willing to assume that the effect sizes are normally distributed, MM is often the method of choice. The method of moments does not depend on any assumptions about the distribution of the random effects, so it has a robustness characteristic that the two other methods (which involve the assumption that the random effects have a normal distribution) do not have. If we are willing to assume a normal distribution of effects, then statisticians tend to prefer ML or REML, which are more efficient than MM (the estimates have smaller variance).

Between MM and REML, ML tends to yield a more precise estimate of T^2 (but with a bias) while REML tends to yield a less biased estimate (but with less precision). With small numbers of studies imprecision can be more important than bias, and so some prefer ML. With more studies, the balance may shift in favor of REML.

ONE-SIDED VS. TWO-SIDED TESTS

Some statistical tests may be performed as either one-sided or two-sided.

- Two-sided tests are appropriate when an effect in either direction would be meaningful.
- One-sided tests are appropriate when we only need to identify an effect in one direction, and an effect in the other direction would have the same implications as zero effect.

In the overwhelming majority of social-science and medical research, we should use a two-tailed test. This is because a two-tailed test allows us to interpret a significant p-value in either direction (favors treated or favors control) whereas a one-tailed test only allows us to interpret a significant p-value in the expected direction.

The reason we use a two-tailed test is that while we typically expect the effect to fall in a specific direction, an effect that was statistically significant in the other direction would still be important. If we expected the treatment to improve survival but it turned out to hurt survival, this would be critically important information. However, if the test had been performed as one-tailed then an effect in the reverse direction (that the treatment is harmful) cannot be statistically significant by definition, even if the computed p-value is < 0.0001. Therefore, except in rare instances, the two tailed test is appropriate.

In the event that you select a one-tailed test, this applies only to the *p*-values for individual covariates on the main results screen.

- It does not affect the confidence interval since this is displayed for lower and upper limits.
- It does not affect the *p*-value for the test of the model. Since this is based on *Q* (or *F*) no direction can be specified and it must be two-tailed.
- It does not affect the *p*-value for a *set* of covariates. Since this is based on *Q* (or *F*) no direction can be specified and it must be two-tailed. (For consistency, this applies even if the set includes only one covariate.)
- It does not affect the *p*-value for a test of the increment. Since this is based on *Q* (or *F*) no direction can be specified and it must be two-tailed.
- It does not affect the confidence interval nor the prediction interval on the plot. Since these are shown for both the lower and upper limit, they are displayed using multipliers for a two-tailed test.

KNAPP-HARTUNG VS. *Z*

Every estimate of a parameter is accompanied by the standard error of that estimate. This affects the confidence interval, the test of significance, and the prediction interval. The standard error can be computed using either the Z option or the Knapp-Hartung option.

In *primary studies*, when we perform a significance test to compare two groups (treated vs. control) we have the option to use either the *Z -test* or the *t-test*. We use the *Z* -test when the population variance is known, and we use the *t*-test when we are using the sample variance to estimate the population variance. The same idea applies to cases where we compare more than two groups. Here, the choice is between chi-squared (when the variance is known) and the *F* statistic (when the variance is estimated). These situations are shown in Table 1.

Table 1 – Test statistics in primary studies

	Variance known	Variance estimated
Two groups	*Z*	*T*
More than two groups	X^2	*F*

The *t*-test is more conservative than the *Z*-test, and the *F*-test is more conservative than the X^2 test. By "more conservative" we mean that (a) the result is less likely to be statistically significant, (b) the confidence interval will be wider. The difference between *Z* and *t* tends to be substantial when the sample size is small (say, less than 30), but declines as the sample size increases. The same holds true for the difference between *F* and X^2.

We are faced with a similar situation in meta-analysis. Since the variances are often being estimated from the observed data, it would make sense to use the *t* distribution to test the null hypothesis and to construct confidence intervals. In fact, though, researchers have traditionally used the *Z* distribution for these purposes.

In the case of a fixed-effect model this distinction turns out to have little practical impact. The only source of error is the variance *within* studies. If studies shared a common error variance (*V*), the error variance for the mean would be V/N. Since the *N* accumulated across studies is typically well over thirty, the difference between *Z* and *t* is usually very small.

However, in the case of a random-effects model, the situation is more complicated. Recall that the error component incorporates two distinct elements – the within-study error and the between-study error. If studies shared a common error variance (*V*), the error variance for the mean would be

$$V_M = \frac{V}{N} + \frac{T^2}{k} \tag{1.81}$$

where N is the number of subjects accumulated across studies, and k is the number of studies. We can justify using *Z* for the within-study error since the N accumulated across studies is typically well over thirty. However, the between-study variance is based on the number of *studies*, which is typically small, and the difference between *Z* and *t* for this component of the variance may be substantial.

The solution proposed by Knapp and Hartung is to address each component of the variance separately. Specifically, we would use the *Z* (or chi-squared) distribution for the within-study variance and the *t* (or *F*) distribution for the between-study variance.

Figure 41 shows a regression using Z, and Figure 214 show the same analysis using Knapp-Hartung. Then, we compare the two.

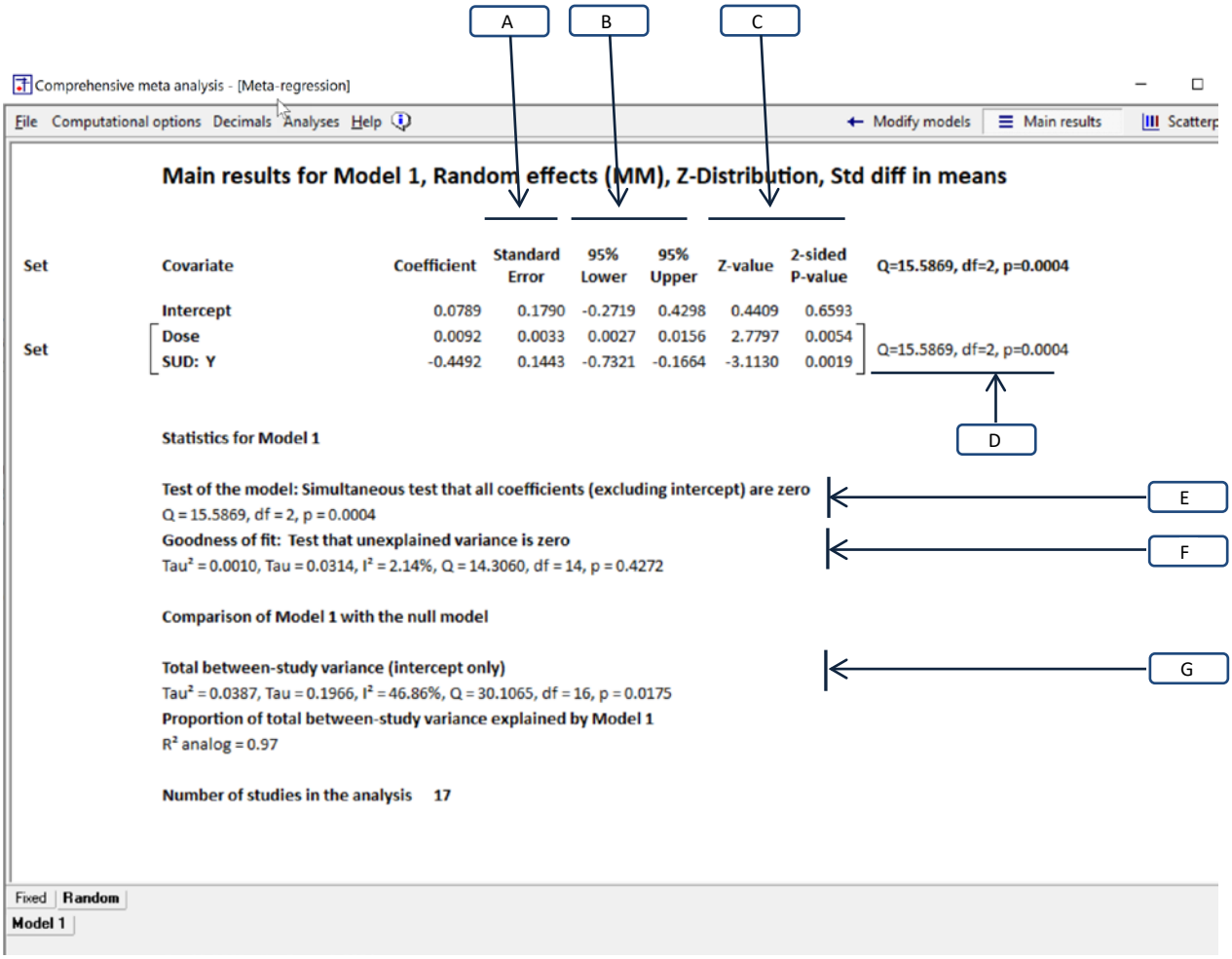


Figure 58 | Main results | Z-Distribution

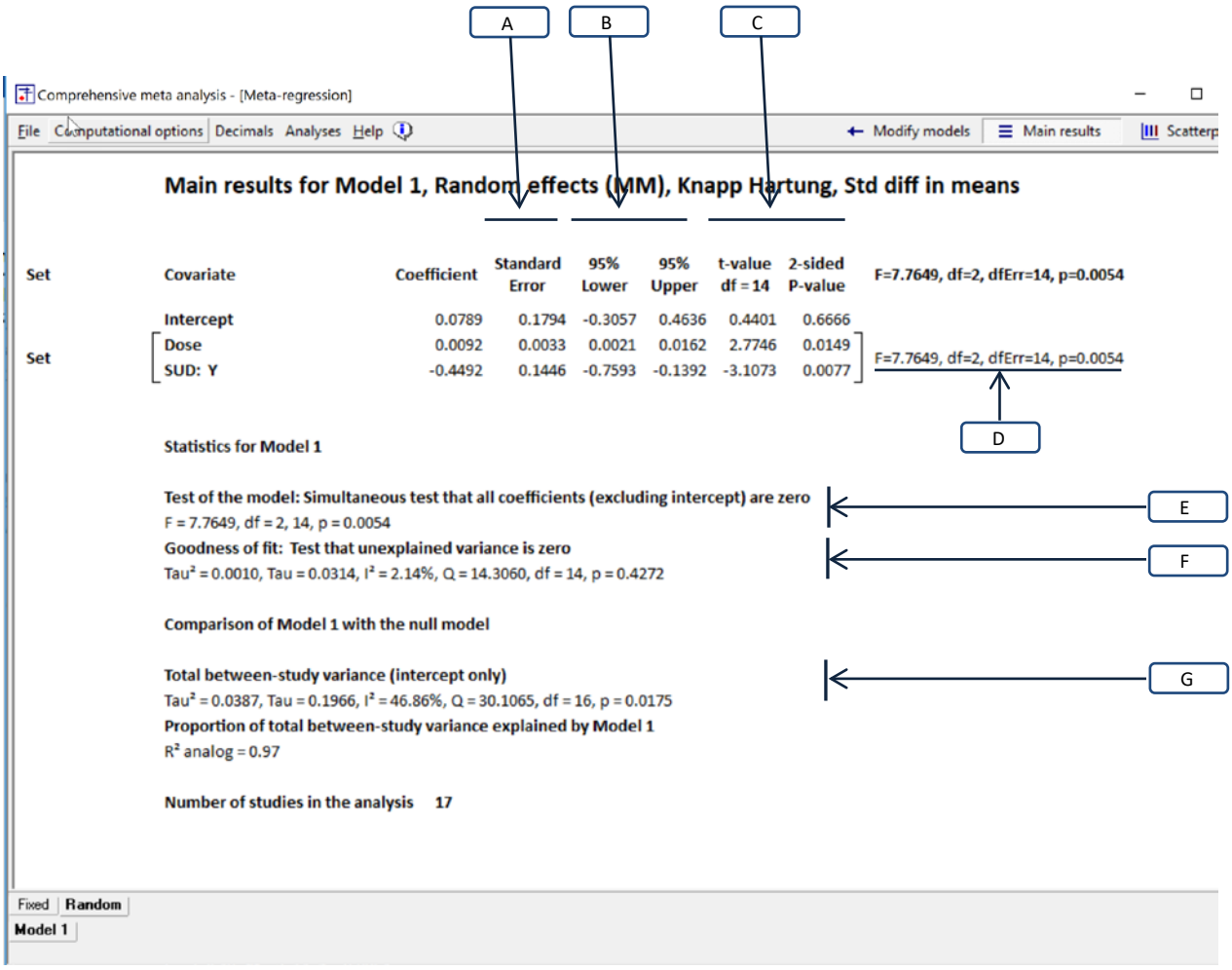


Figure 59 | Main results | Knapp-Hartung

Table of coefficients

When we move from a *Z*-score to Knapp Hartung

- The coefficients do not change
- The standard error increases [A]
- The confidence interval width increases [B]
- The *Z*-score is replaced by a (smaller) *t*-score [C]
- The *p*-value becomes larger (less significant) [C]
- The *Q*-value for a set is replaced by a (smaller) *F*-score [D]
- The *p*-value for a set becomes less significant [D]

Test of the model

When we move from a *Z*-score to Knapp Hartung

- The *Q*-value is replaced by a (smaller) *F*-value
- The *p*-value becomes less significant

Goodness of fit

When we move from a *Z*-score to Knapp Hartung [F]

- The numbers do not change. This is because the Knapp-Hartung adjustment only applies to the T^2 part of the variance, but the goodness of fit test is computed assuming $T^2$is zero.

Comparison of Model 1 with the null model

When we move from a *Z*-score to Knapp Hartung [G]

- The numbers do not change. This is because this comparison employs weights based on within-study variance (*V*), and the Knapp Hartung adjustment only affects between-study variance (T^2).

The same differences apply also on other screens. Of note, the lines for the confidence interval and the prediction interval in the scatterplot will be wider when Knapp-Hartung is in effect.

While it is always true that the *p*-value will be the same or higher (further from zero) for Knapp-Hartung (KH), the extent of the difference depends on the amount of between-study variance and the number of studies. To the extent that the between-study population variance is small and/or the number of studies is large, the between-study error variance will be small, and the difference between the *Z* option and the KH option will tend to be relatively small. Conversely, to the extent that the between-study population variance is large and/or the number of studies is low, the difference between the two options will tend to be relatively large.

You do not need to return to the [Modify models] screen to switch between *Z* and Knapp-Hartung. Rather, if you're already looking at the results you can simply change the setting and the results will change.

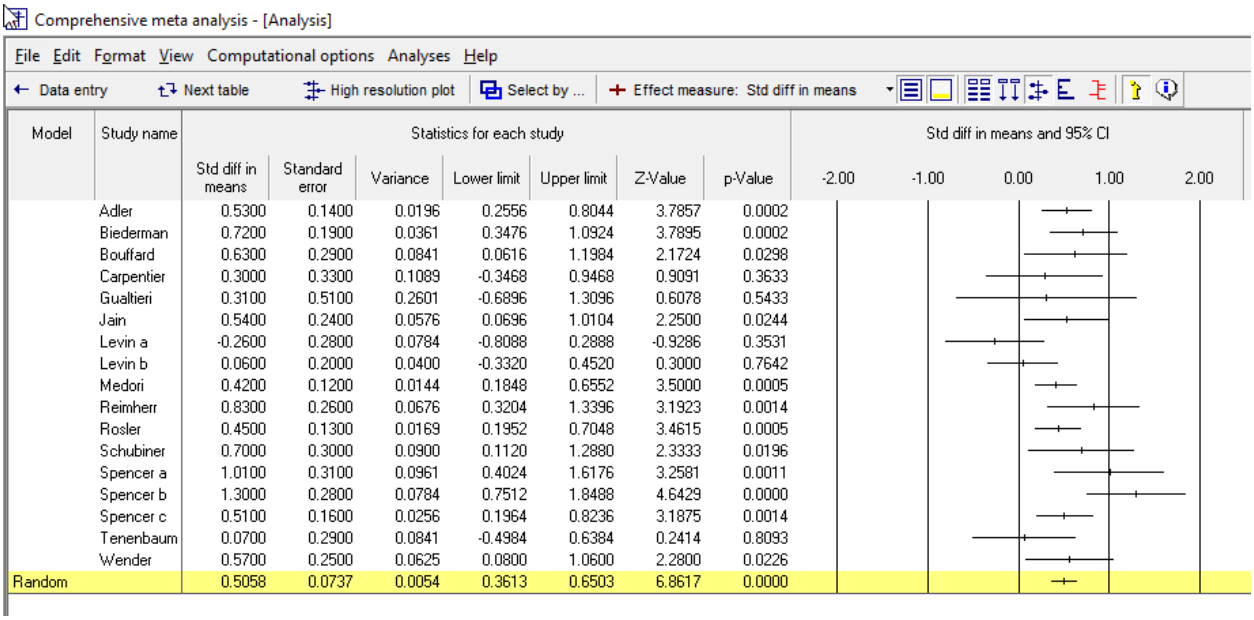
Since the Knapp-Hartung option is intended to address uncertainty in between-studies variance, most people who use it do so only for the random-effects model. In CMA, the Knapp-Hartung option is only available for random-effects models.

The intent of the Knapp-Hartung adjustment is to improve the accuracy of *p*-values, confidence intervals, and prediction intervals. Higgins and Thompson (2004) proposed an approach that bypasses the sampling distributions and instead employs a permutation test to yield a *p*-value. Using this approach we would compute the *Z*-score corresponding to the observed covariate. Then, we would randomly redistribute the covariates among studies and see what proportion of these re-distributions yield a *Z*-score exceeding the one that we had obtained. This proportion may be viewed as an exact *p*-value. This option is not implemented in CMA.

Using the Knapp-Hartung adjustment for a simple analysis or for subgroups

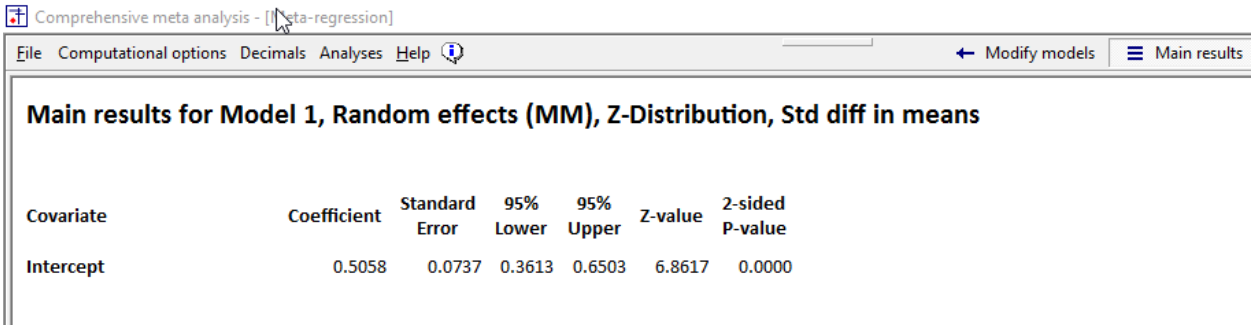
While these adjustments can be applied to any use of the random-effects model (that is, for a single group of studies, for a subgroup analysis, and for meta-regression), to date we have only implemented them for the meta-regression. We plan to update the other modules in the future. In the interim, one could use the regression module to obtain Knapp-Hartung estimates for a simple analysis or for a subgroups analysis by using the regression module with no covariates or with one categorical covariate.

In the ADHD example, using random-effects weights the simple analysis yields an effect size of 0.5058 with a standard error of 0.0737, a confidence interval of 0.3613 to 0.6503. The test of null hypothesis that the mean effect size is zero yields a Z-value of 6.8617 and a p-value of < 0.0001.

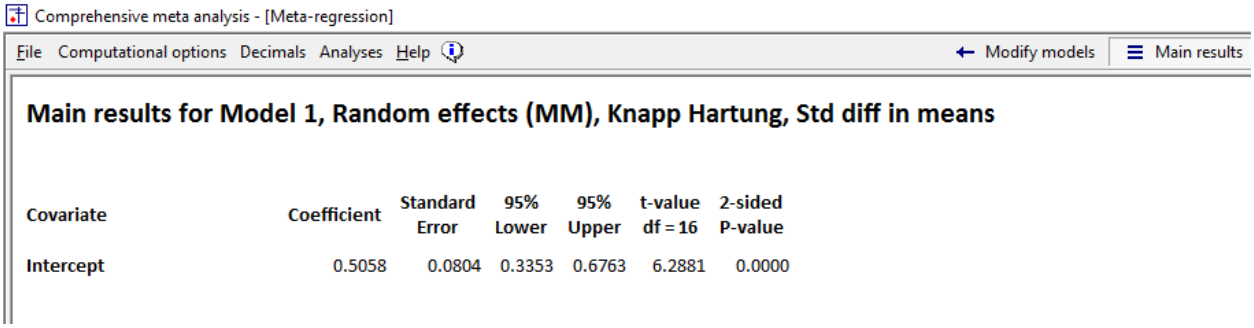


Model	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Adler	0.5300	0.1400	0.0196	0.2556	0.8044	3.7857	0.0002
	Biederman	0.7200	0.1900	0.0361	0.3476	1.0924	3.7895	0.0002
	Bouffard	0.6300	0.2900	0.0841	0.0616	1.1984	2.1724	0.0298
	Carpentier	0.3000	0.3300	0.1089	-0.3468	0.9468	0.9091	0.3633
	Gualtieri	0.3100	0.5100	0.2601	-0.6896	1.3096	0.6078	0.5433
	Jain	0.5400	0.2400	0.0576	0.0696	1.0104	2.2500	0.0244
	Levin a	-0.2600	0.2800	0.0784	-0.8088	0.2888	-0.9286	0.3531
	Levin b	0.0600	0.2000	0.0400	-0.3320	0.4520	0.3000	0.7642
	Medori	0.4200	0.1200	0.0144	0.1848	0.6552	3.5000	0.0005
	Reimherr	0.8300	0.2600	0.0676	0.3204	1.3396	3.1923	0.0014
	Rosler	0.4500	0.1300	0.0169	0.1952	0.7048	3.4615	0.0005
	Schubiner	0.7000	0.3000	0.0900	0.1120	1.2880	2.3333	0.0196
	Spencer a	1.0100	0.3100	0.0961	0.4024	1.6176	3.2581	0.0011
	Spencer b	1.3000	0.2800	0.0784	0.7512	1.8488	4.6429	0.0000
	Spencer c	0.5100	0.1600	0.0256	0.1964	0.8236	3.1875	0.0014
	Tenenbaum	0.0700	0.2900	0.0841	-0.4984	0.6384	0.2414	0.8093
	Wender	0.5700	0.2500	0.0625	0.0800	1.0600	2.2800	0.0226
Random		0.5058	0.0737	0.0054	0.3613	0.6503	6.8617	0.0000

If we run the regression with no covariates, the predicted effect size is simply the intercept, which is also the mean. If we select the method of moments and the Z-distribution, the statistics for the intercept will be identical to those we saw in the simple analysis. To wit, the intercept is 0.5058 with a standard error of 0.0737, a confidence interval of 0.3613 to 0.6503. The test of null hypothesis that the mean effect size is zero yields a Z-value of 6.8617 and a p-value of < 0.0001.

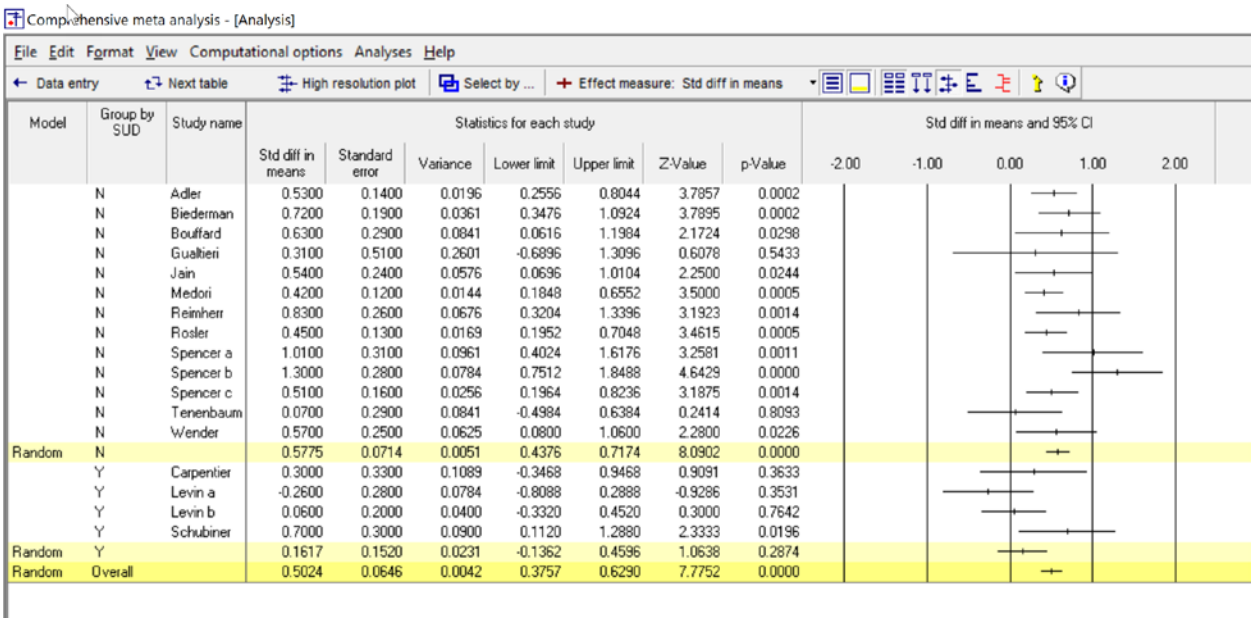


If we select the Knapp-Hartung option, the intercept is 0.5058 with a standard error of 0.0804, a confidence interval of 0.3353 to 0.6763. The test of null hypothesis that the mean effect size is zero yields a t-value of 6.2881 with df=16 and a p-value of < 0.0001. In this example the adjustment is relatively modest but in many analyses (especially when the number of studies is low) the adjustment will be substantial.



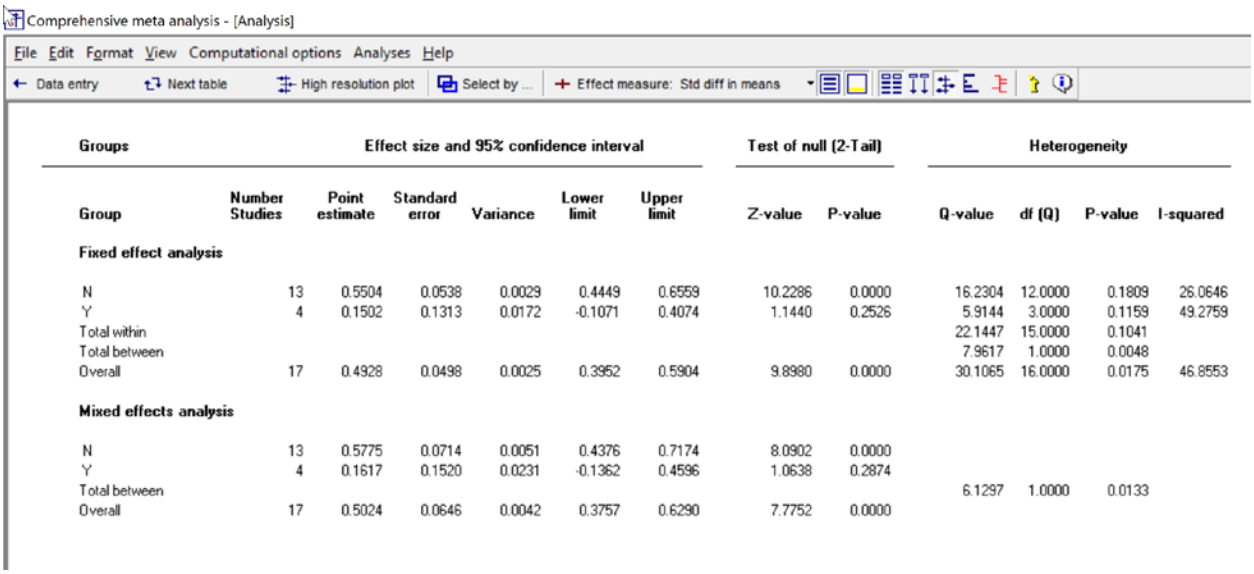
The same applies to subgroups. If we group by subgroups in the main analysis module, the program will display the confidence intervals and p-value based on the Z distribution. If we want to use the Knapp-Hartung adjustment, we can do so by running the analysis as a regression.

This shows an analysis grouped by SUD



Model	Group by SUD	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	N	Adler	0.5300	0.1400	0.0196	0.2556	0.8044	3.7857	0.0002
	N	Biederman	0.7200	0.1900	0.0361	0.3476	1.0924	3.7895	0.0002
	N	Bouffard	0.6300	0.2900	0.0841	0.0616	1.1984	2.1724	0.0298
	N	Gualtieri	0.3100	0.5100	0.2601	-0.6896	1.3096	0.6078	0.5433
	N	Jain	0.5400	0.2400	0.0576	0.0696	1.0104	2.2500	0.0244
	N	Medori	0.4200	0.1200	0.0144	0.1848	0.6552	3.5000	0.0005
	N	Reimherr	0.8300	0.2600	0.0676	0.3204	1.3396	3.1923	0.0014
	N	Rosler	0.4500	0.1300	0.0169	0.1952	0.7048	3.4615	0.0005
	N	Spencer a	1.0100	0.3100	0.0961	0.4024	1.6176	3.2581	0.0011
	N	Spencer b	1.3000	0.2800	0.0784	0.7512	1.8488	4.6429	0.0000
	N	Spencer c	0.5100	0.1600	0.0256	0.1964	0.8236	3.1875	0.0014
	N	Tenenbaum	0.0700	0.2900	0.0841	-0.4984	0.6384	0.2414	0.8093
	N	Wender	0.5700	0.2500	0.0625	0.0800	1.0600	2.2800	0.0226
Random	N		0.5775	0.0714	0.0051	0.4376	0.7174	8.0902	0.0000
	Y	Carpentier	0.3000	0.3300	0.1089	-0.3468	0.9468	0.9091	0.3633
	Y	Levin a	-0.2600	0.2800	0.0784	-0.8088	0.2888	-0.9286	0.3531
	Y	Levin b	0.0600	0.2000	0.0400	-0.3320	0.4520	0.3000	0.7642
	Y	Schubiner	0.7000	0.3000	0.0900	0.1120	1.2880	2.3333	0.0196
Random	Y		0.1617	0.1520	0.0231	-0.1362	0.4596	1.0638	0.2874
Random	Overall		0.5024	0.0646	0.0042	0.3757	0.6290	7.7752	0.0000

Figure 60



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												
N	13	0.5504	0.0538	0.0029	0.4449	0.6559	10.2286	0.0000	16.2304	12.0000	0.1809	26.0646
Y	4	0.1502	0.1313	0.0172	-0.1071	0.4074	1.1440	0.2526	5.9144	3.0000	0.1159	49.2759
Total within									22.1447	15.0000	0.1041	
Total between									7.9617	1.0000	0.0048	
Overall	17	0.4928	0.0498	0.0025	0.3952	0.5904	9.8980	0.0000	30.1065	16.0000	0.0175	46.8553
Mixed effects analysis												
N	13	0.5775	0.0714	0.0051	0.4376	0.7174	8.0902	0.0000				
Y	4	0.1617	0.1520	0.0231	-0.1362	0.4596	1.0638	0.2874				
Total between									6.1297	1.0000	0.0133	
Overall	17	0.5024	0.0646	0.0042	0.3757	0.6290	7.7752	0.0000				

Figure 61

To run the same analysis using regression we cannot simply run the regression and include SUD as a covariate, as shown in

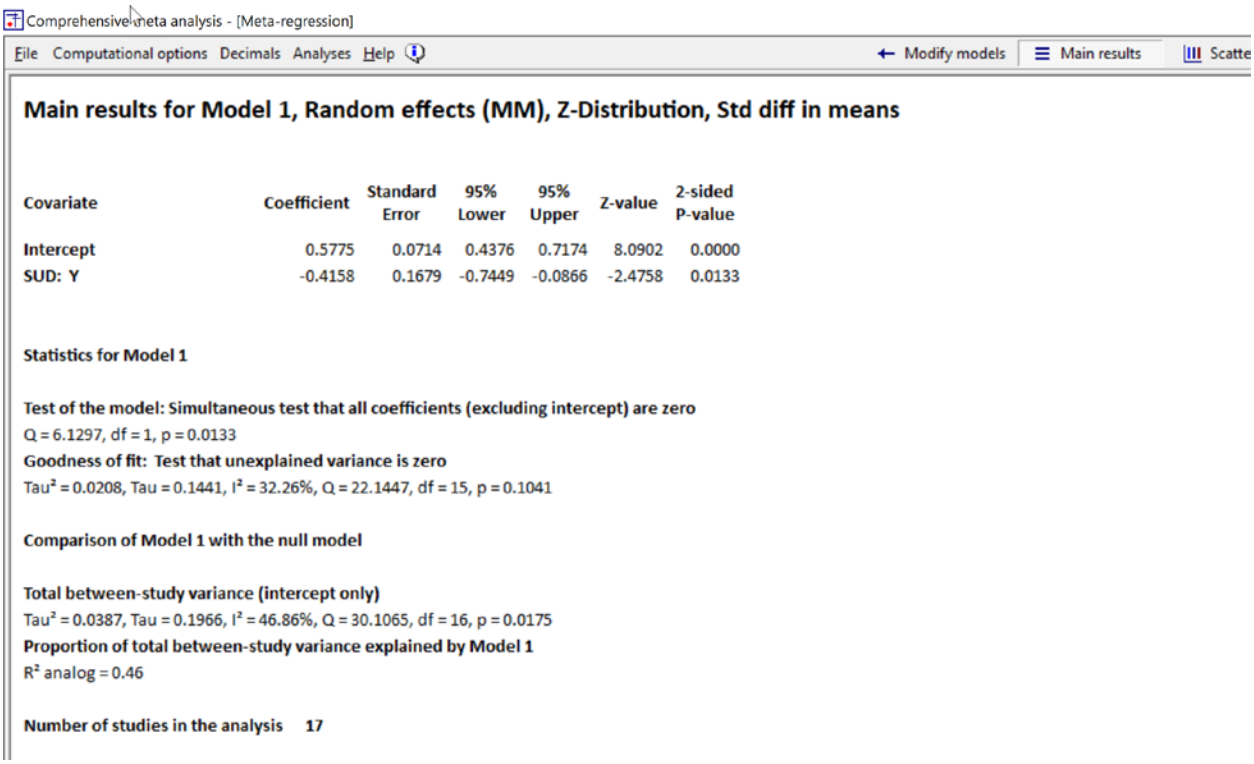


Figure 62

If we did that, the line for the intercept would give us the mean for SUD | N with the corresponding standard error. However, the line for SUD | Y would *not* give us the mean for SUD | Y. Rather, it would give is the difference in means between SUD | N and SUD | Y. And, the standard error would not be the standard error of the mean for SUD | Y. Rather, it would be the standard error of the difference. Concretely, the error variance for SUD | N is 0.005095 and the error variance for SUD | Y is 0.023105. The error variance of the difference is the sum of the two, 0.028200. And the standard error of the difference is 0.1679. This is the value displayed as the SE for SUD |Y.

Rather, if we want to display the mean for each group we need to run the analysis without the intercept. For details, see "When to omit the intercept". The results of this analysis are shown in Figure 63. Now, the two rows display the means for the two groups along with the standard error, confidence interval, and p-value for the means. The values displayed here are identical to the values displayed in Figure 61.

Now that we have this set up as a regression, we can change the computational option to Knapp-Harting. The results are shown in Figure 64.

For SUD | N, the mean has not changed but the standard error is larger, the confidence interval is wider, the Z value is smaller, and the p-value is further from zero (though it still displays as 0 to four decimal places).

For SUD | Y, the mean has not changed but the standard error is larger, the confidence interval is wider, the Z value is smaller, and the p-value is further from zero.

Note that the values have not changed dramatically, even for SUD | Y, where there are only four studies. This is because the value of T2 is based on all 17 studies in the analysis, not just the four in this subgroup. If we had run an analysis based on the four studies alone, the difference between using Z and Knapp-Hartung would have been substantially larger.

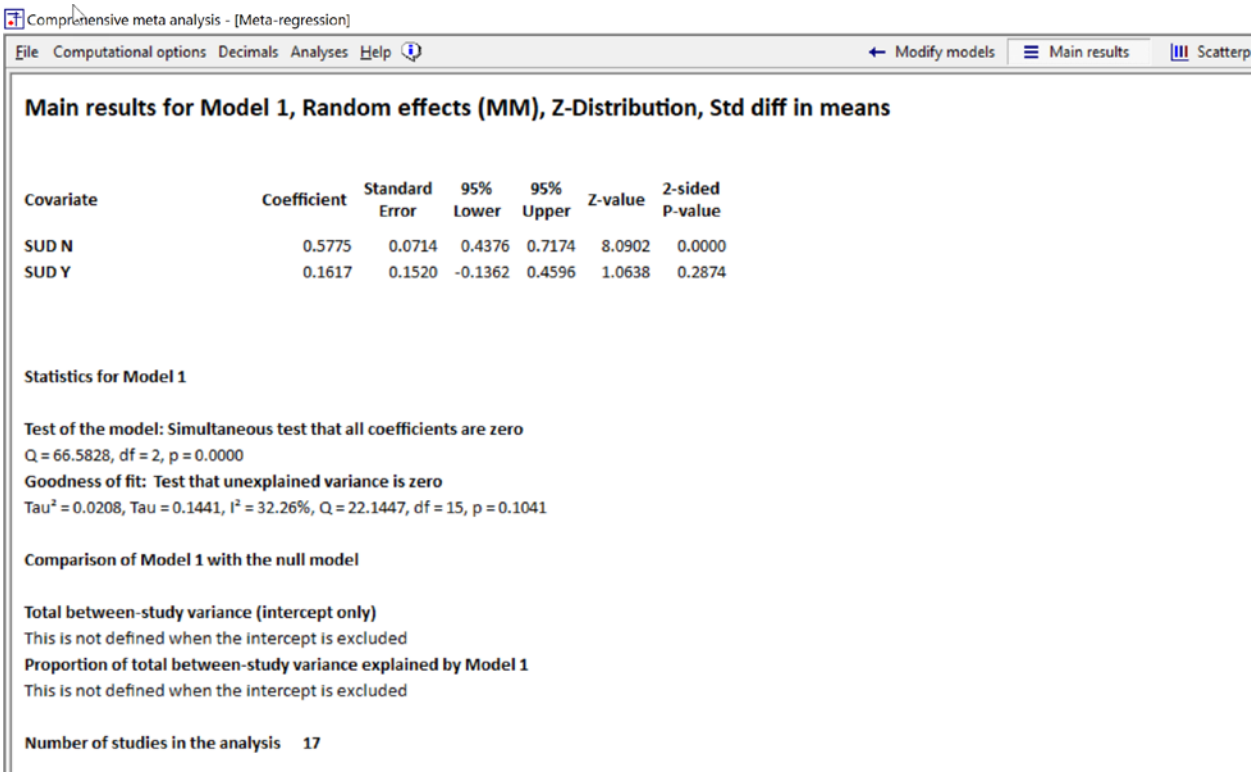


Figure 63

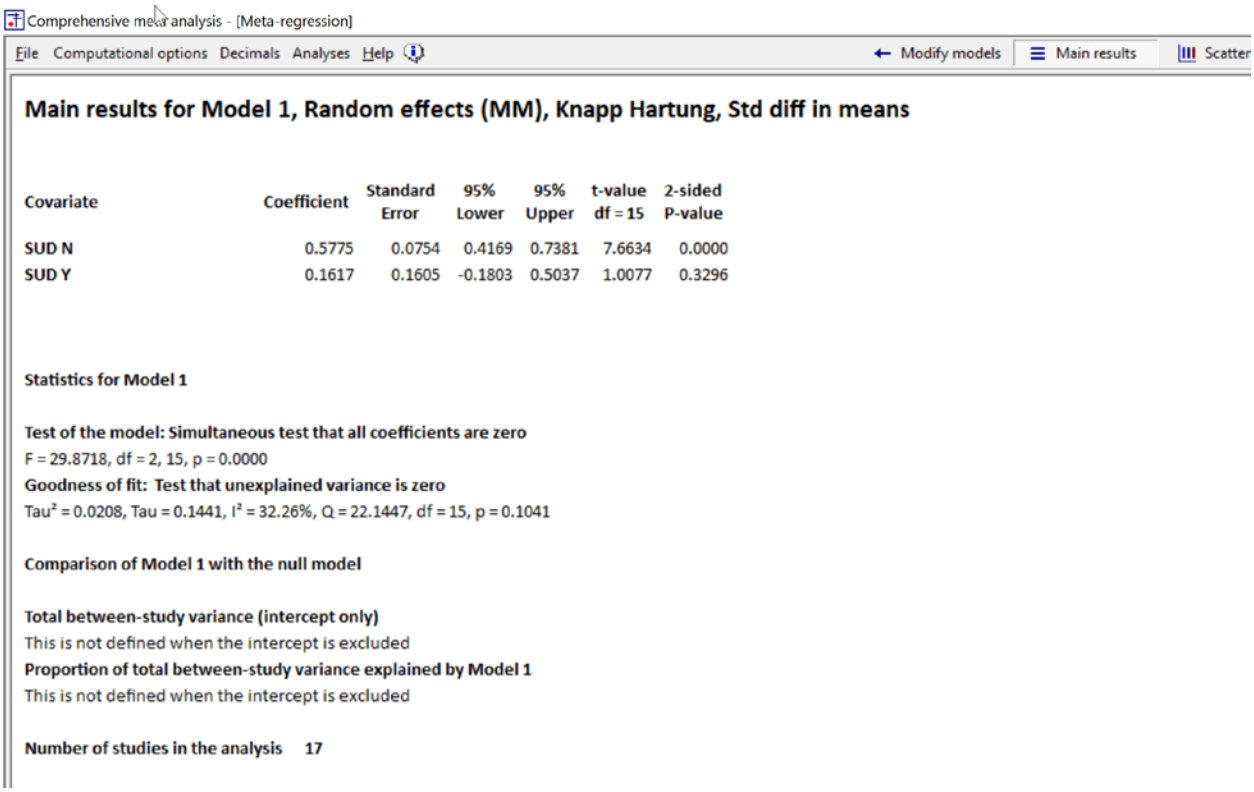

Comprehensive meta analysis - [Meta-regression]
File Computational options Decimals Analyses Help
Modify models
Main results
Scatter
Main results for Model 1, Random effects (MM), Knapp Hartung, Std diff in means
Covariate
Coefficient
Standard Error
95% Lower
95% Upper
t-value df = 15
2-sided P-value
SUD N 0.5775 0.0754 0.4169 0.7381 7.6634 0.0000
SUD Y 0.1617 0.1605 -0.1803 0.5037 1.0077 0.3296
Statistics for Model 1
Test of the model: Simultaneous test that all coefficients are zero
F = 29.8718, df = 2, 15, p = 0.0000
Goodness of fit: Test that unexplained variance is zero
Tau² = 0.0208, Tau = 0.1441, I² = 32.26%, Q = 22.1447, df = 15, p = 0.1041
Comparison of Model 1 with the null model
Total between-study variance (intercept only)
This is not defined when the intercept is excluded
Proportion of total between-study variance explained by Model 1
This is not defined when the intercept is excluded
Number of studies in the analysis 17

Figure 64

Assumptions for the use of the Knapp-Hartung Adjustment

The KH adjustment assumes that the standardized residuals are identically distributed. If there are some outliers, this assumption will be violated and the adjustment might not work. In particular, if some effects have a very small standard error, the standardized residual for these effects would be high. This is especially a concern when the effect size index is the prevalence in one group, and the prevalence for some studies is near zero. If the analysis is based on the logit transformation, these studies will have a small standard error and a large distance from the mean, yielding a potentially large standardized residual.

The Knapp-Hartung variance takes the theoretical variance and inflates it by a factor of QR = SUM{wi*(Ti – Tbar)2}, where wi* is the random effect weight.

Each term of the sum is the square of a standardized residual (Ti – Tbar)/sqrt(vi*) [Here I mean standardized as they use it in regression diagnostics—they do not have unit variance, but variance a little less than 1).

In any event, KH assumes identically distributed standardized residuals, so if they aren't (e.g., if there are some outliers), then KH may not give valid results.

ONE-POINT OR SIMULTANEOUS INTERVALS

The program allows you to plot the confidence interval for the regression line. For any value of dose, the *mean* effect size may not fall precisely on the regression line but in most cases it will fall within the range indicated by the confidence interval. Similarly, the program allows you to plot the prediction interval for the regression line. For any value of dose, the true effect size for any single study will usually fall in the range indicated by the prediction interval.

There are two useful options for plotting these intervals. We can plot an interval that is accurate for *any single point* on the graph, or an interval that is accurate for *all points* on the graph simultaneously. For the second option the program uses a multiplier based on the Sheffé adjustment to widen the confidence interval.

Click Computational options > and select one of the following

- One point. If we were to randomly select *a single dose*, in 95% of all possible regressions, the mean effect size *for that dose* would fall within the confidence interval. And the true effect size for 95% of all studies *at that dose* would fall within the prediction interval.

- Simultaneous. If we were to look at *all doses*, in 95% of all possible regressions, the mean effect size would fall within the confidence interval. And the true effect size for 95% of all studies *at any dose* would fall within the prediction interval.

When we choose the simultaneous option the lines will move further from the regression line.

How to compute CI for any point
Why the lines are concave

CONFIDENCE LEVEL

By default, the program displays the 95% interval, but you can choose to display many other intervals as well. These range from the 99% interval to the 68% interval.

To set the confidence level to 95%, click Computational options > CI Level > 95%

This setting affects the following

- The confidence interval displayed on the main results screen
- The confidence interval on the plot
- The prediction interval on the plot

MISTAKES TO AVOID WHEN REPORTING A META-REGRESSION

It's not unusual to see reports of an analysis that include a p-value, or even a covariate, but leave the reader with no idea of what's going on. There's a limited amount of space in an abstract (and even in a full paper). It's more important to deliver the key take-home message. In reporting on the relationship between Dose and effect size

Mistake – give the p-value but not the direction and magnitude of the effect

What we do want to emphasize about this report is that we focus on the magnitude of relationships and the clinical implications of the findings. For example, it's not enough to say that the effect size is related to the dose – it's also important to mention what the effect size is at the lowest and the highest dose. It's also critically important to be transparent about the limitations of the regression. We say that a higher dose is associated with a larger effect size (not that it causes a larger effect size). Then we say explicitly that the relationship may not be causal, and we explain why. In the case of SUD we follow the excellent example set by the paper's authors, and point out one potentially important confound.

Treat as causal

Base on small number of studies (Formulation)

Wrong model

Example of wrong and better for each one

NEED

CATEGORICAL COVARIATES

Categorical covariates are covariates that represent a category or group, rather than a numerical score. For example, the covariate SUD reflects the fact that the study either included or excluded patients with substance-abuse disorders. Each study is coded as "N" (excluded SUD) or "Y" (included SUD).

CHANGE TO SUD

When we perform a subgroups analysis (as with an analysis of variance in a primary study) we can work directly with categorical covariate and classify each study as "N" or "Y". By contrast, for a regression we need to work with numbers, not labels. Therefore, rather than working with the original variable we create so-called "dummy variables", numeric variables that stand for a group or category.

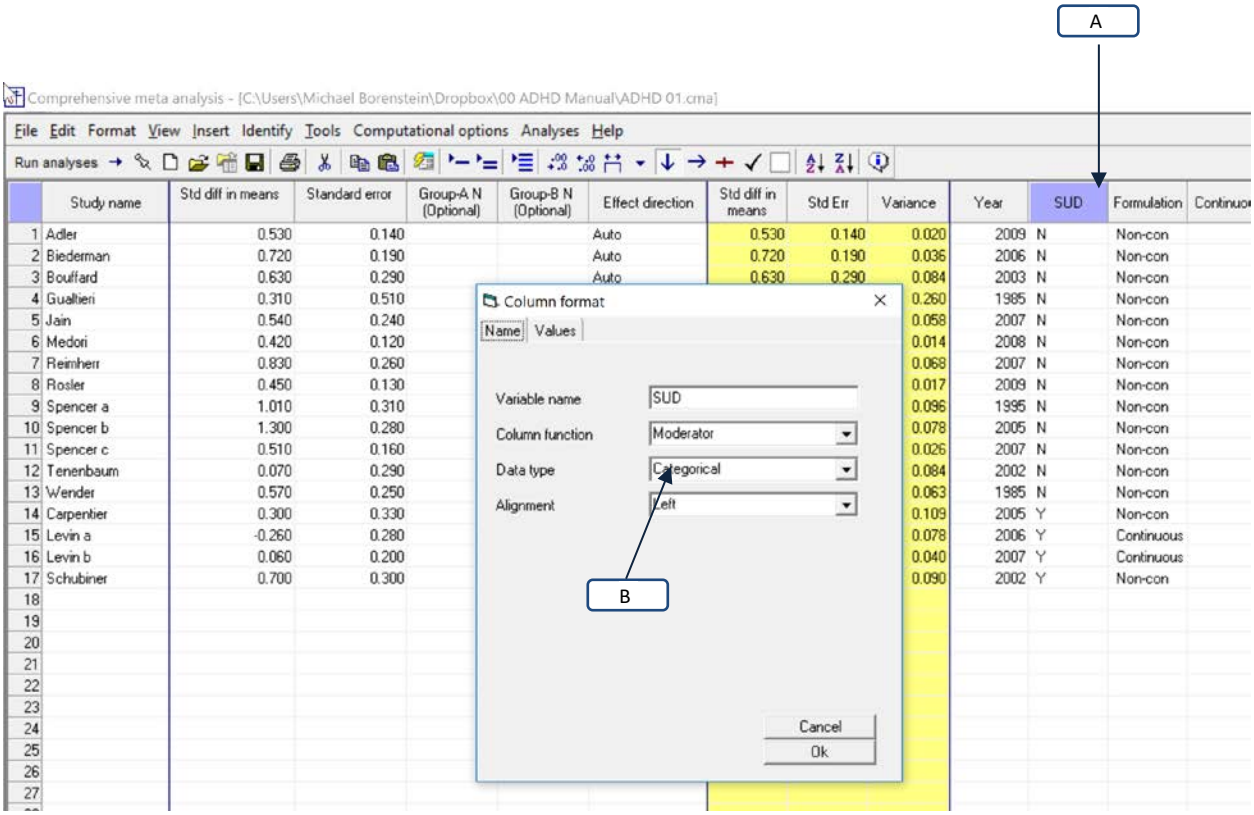
The general rule is that a variable with *m* categories will be represented by *m*-1 dummy-variables. For example, if we have a categorical variable called Gender which is coded either Male or Female, we would not create two dummy variables, but rather one. We could call this Male and code it 0 for Female and 1 for Male. The reason that we don't create a second variable (Female) is that this would contain precisely the same information as the first variable.

While it's possible to create dummy variables manually, the program will also create them automatically, which is generally the simpler approach. Concretely, when you enter variables into the analysis, the program will recognize which ones have been designated as categorical. Then it automatically create one (or more) dummy variables, assign a code for each study, and enter these dummy-variables into the analysis.

Below, we present two examples. The first is for a categorical variable with two groups. The second is for a categorical variable with three groups. From there, it's a simple extension to any number of groups.

Dummy variable for a covariate with two groups

The dataset includes a variable called SUD [A], which is coded N (excludes SUD) or Y (includes SUD). Note that Formulation must be defined as a Moderator variable > Categorical [B].



	Study name	Std diff in means	Standard error	Group-A N (Optional)	Group-B N (Optional)	Effect direction	Std diff in means	Std Err	Variance	Year	SUD	Formulation	Continuo
1	Adler	0.530	0.140			Auto	0.530	0.140	0.020	2009	N	Non-con	
2	Biederman	0.720	0.190			Auto	0.720	0.190	0.036	2006	N	Non-con	
3	Bouffard	0.630	0.290			Auto	0.630	0.290	0.084	2003	N	Non-con	
4	Gualtieri	0.310	0.510						0.260	1985	N	Non-con	
5	Jain	0.540	0.240						0.058	2007	N	Non-con	
6	Medori	0.420	0.120						0.014	2008	N	Non-con	
7	Reimherr	0.830	0.260						0.068	2007	N	Non-con	
8	Rosler	0.450	0.130						0.017	2009	N	Non-con	
9	Spencer a	1.010	0.310						0.096	1995	N	Non-con	
10	Spencer b	1.300	0.280						0.078	2005	N	Non-con	
11	Spencer c	0.510	0.160						0.026	2007	N	Non-con	
12	Tenenbaum	0.070	0.290						0.084	2002	N	Non-con	
13	Wender	0.570	0.250						0.063	1985	N	Non-con	
14	Carpentier	0.300	0.330						0.109	2005	Y	Non-con	
15	Levin a	-0.260	0.280						0.078	2006	Y	Continuous	
16	Levin b	0.060	0.200						0.040	2007	Y	Continuous	
17	Schubiner	0.700	0.300						0.090	2002	Y	Non-con	
18													
19													
20													
21													
22													
23													
24													
25													
26													
27													

CMA is able to create a dummy variable for *SUD* automatically.

In the regression module (Figure 65),

- Click on Show Covariates [A]
- Click on SUD [B]
- Click on Edit reference group [C]
- Select [N] [D]
- Click [Add to main screen] [E]

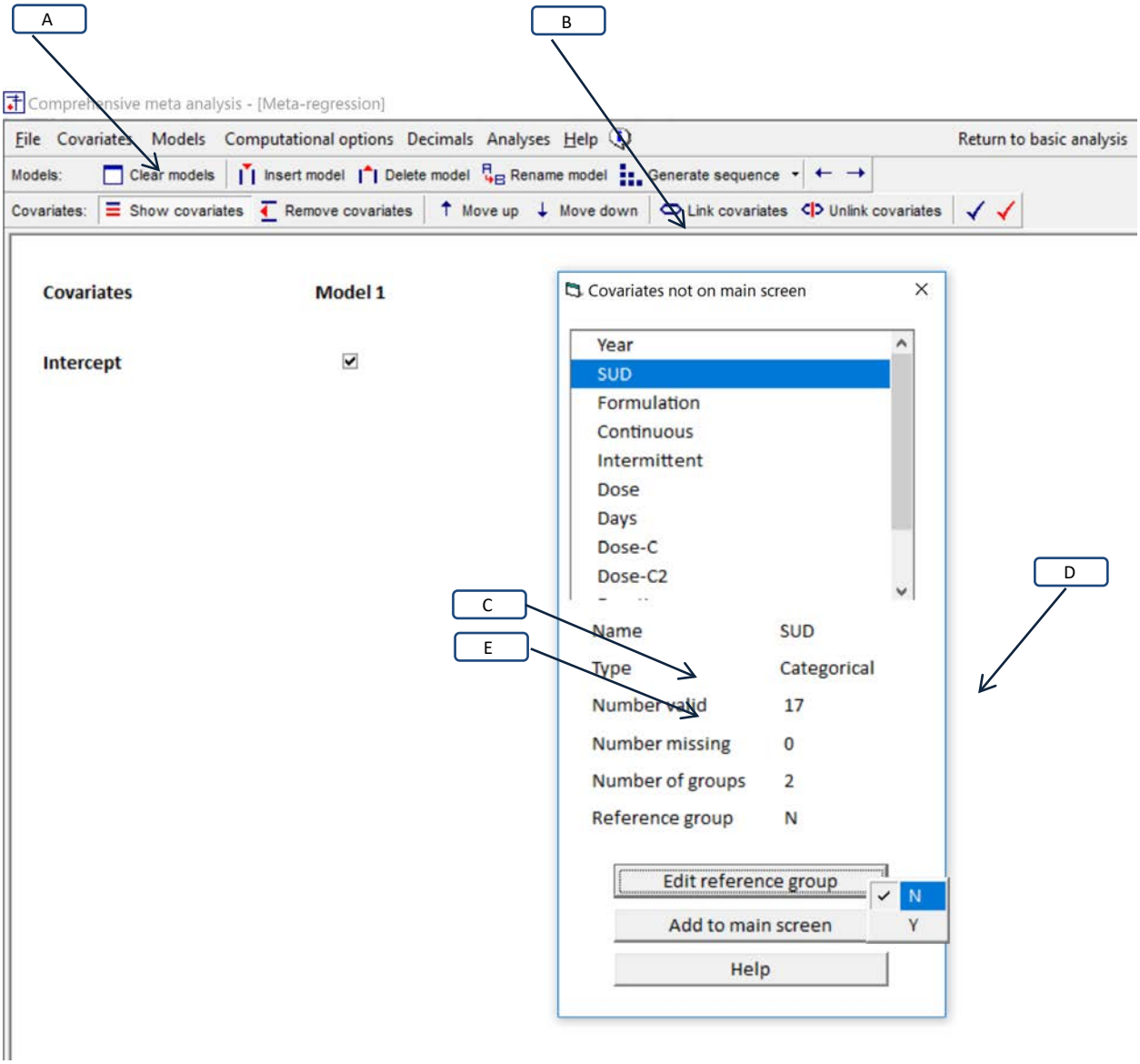


Figure 65 | Creating dummy variables

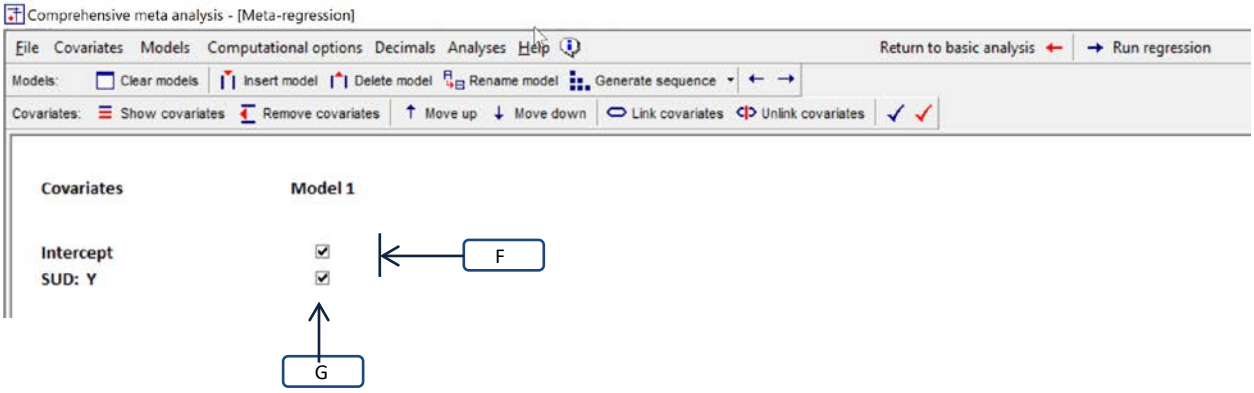


Figure 66 | Creating dummy variables

Since there are two (*m*) groups, there is one (*m*-1) covariate.

The dummy variable is "SUD: Y". Following the conventions proposed by Cohen, studies are coded "1" if they belong to the dummy-variable's group name. A study is coded 1 for "SUD" if the SUD patients are *included*, or 0 if they are excluded.

A positive coefficient will mean that the studies which included SUD patients had a higher effect size than the reference group, while a negative coefficient will mean that the studies which included SUD patients had a lower effect size than the reference group.

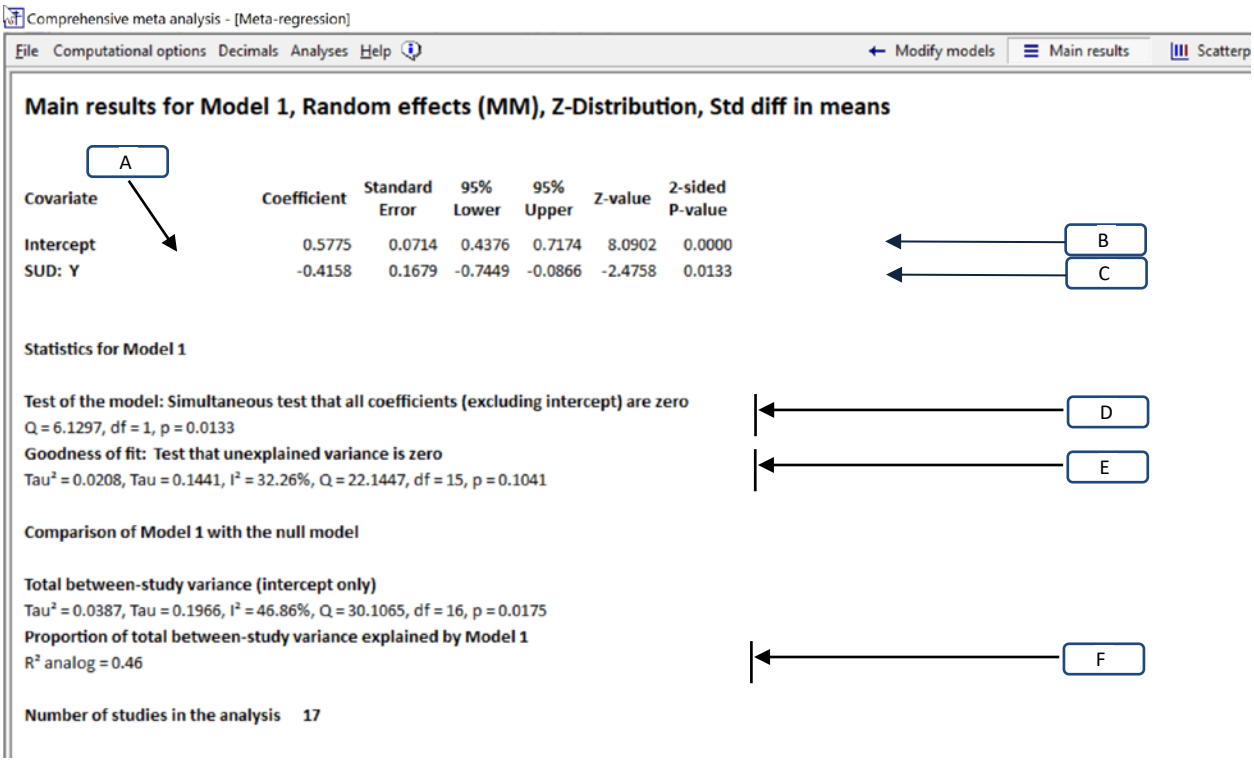


Figure 67 | Dummy variables | SUD with "N" as the reference group

The coefficient for SUD: Y is -0.4158. This tells us that the mean effect size for the studies that included SUD patients was 0.4157 points lower than the mean for studies that excluded these patients. The 95% confidence interval for the difference in means is -0.7449 to -0.0866. The Z-value for the difference is -2.4758 and the p-value is 0.0133. We conclude that the effect size is probably lower in studies that enrolled SUD patients as compared with studies that excluded these patients.

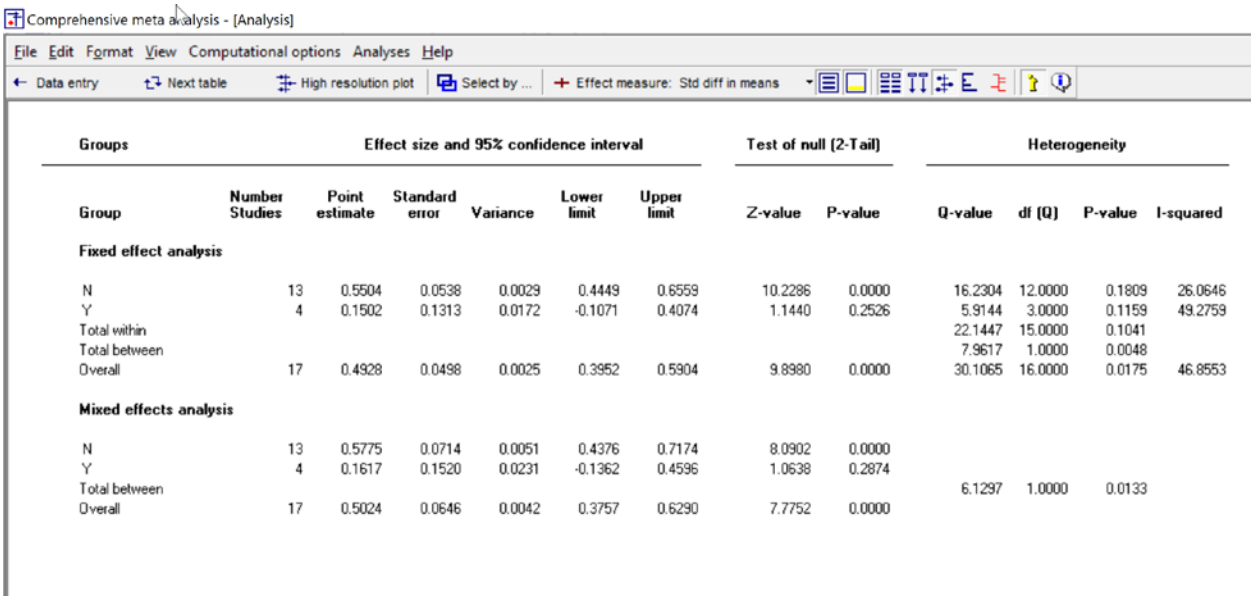
Since there is only one covariate in this analysis, the test of the model is identical to the test of the covariate. For the model the Q-value is 6.1297 with 1 degree of freedom and p=0.0133.

T2, the variance of true effects about the subgroup means, is 0.0208. T, the standard deviation of true effects about the subgroup means, is 0.1441. I2, the proportion of variance in observed effects that reflects variation in true effects rather than sampling error, is 32.26%. The Q-value for the residual variance is 22.1447 with 15 degrees of freedom and p=0.1041. The criterion alpha for this test is 0.10, so this skirts statistical significance. If we treat it as statistically significant we would say there is evidence that the true effects vary within subgroups.

The R2 analog is 0.46, which tells us that some 46% of the variance in true effects can be explained by the difference between subgroups.

To this point we've focused on the difference between groups, but it may also be important to know the absolute effect size in each group. For the SUD| N group the mean effect is simply the intercept, which is 0.5775. For the SUD | Y group the mean effect is given by the computation 0.5775 − 0.4158 or 0.1617.

We could have obtained the same results using a subgroups analysis as shown in Figure 68, where the mean for the SUD | N group is 0.5775, the mean for the SUD | Y group is 0.1617, and the difference (which we compute as 0.4168) yields a Q-value of 6.1297 with 1 degree of freedom and p=0.0133.



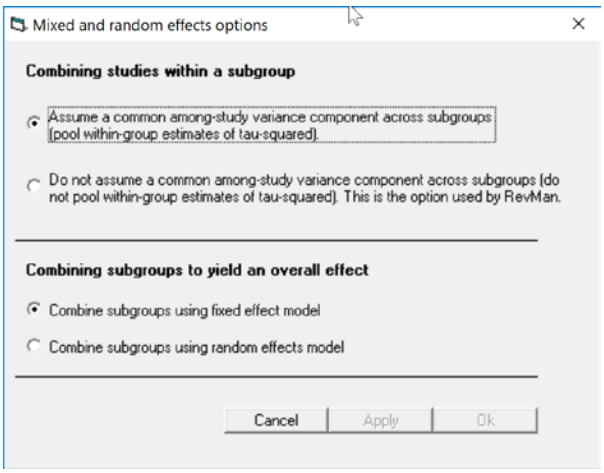
Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
N	13	0.5504	0.0538	0.0029	0.4449	0.6559	10.2286	0.0000	16.2304	12.0000	0.1809	26.0646
Y	4	0.1502	0.1313	0.0172	-0.1071	0.4074	1.1440	0.2526	5.9144	3.0000	0.1159	49.2759
Total within									22.1447	15.0000	0.1041	
Total between									7.9617	1.0000	0.0048	
Overall	17	0.4928	0.0498	0.0025	0.3952	0.5904	9.8980	0.0000	30.1065	16.0000	0.0175	46.8553
Mixed effects analysis												
N	13	0.5775	0.0714	0.0051	0.4376	0.7174	8.0902	0.0000				
Y	4	0.1617	0.1520	0.0231	-0.1362	0.4596	1.0638	0.2874				
Total between									6.1297	1.0000	0.0133	
Overall	17	0.5024	0.0646	0.0042	0.3757	0.6290	7.7752	0.0000				

Figure 68

However, the regression allows us to include an additional covariate. In that case, the regression will give us the unique impact of Formulation when the other covariate is partialled.

Note. For the subgroups analysis, Computational options > Random and mixed-effect options must be set as shown here.

- At the top, select "Assume a common among-study variance component across subgroups (pool within-group estimates of tau-squared)."
- At the bottom, select "Combine subgroups using fixed-effect model". *This option should be selected even if using random-effects.* For a discussion of these options, see chapter _____,

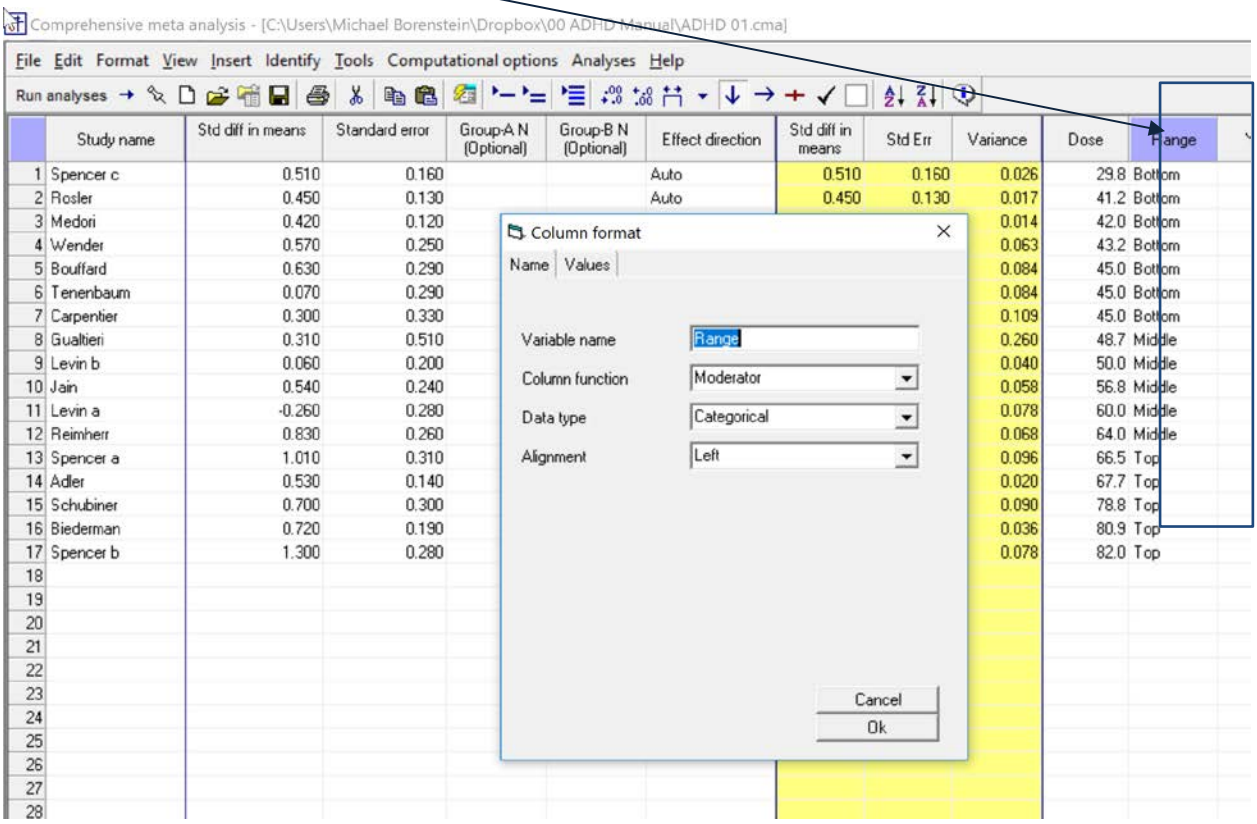


Dummy variables for a covariate with three or more groups

The next example shows how to work with a categorical variable that has three groups. The dataset does not include any variables with three categories, so we'll create one for the purpose of this example.

One of the variables in our data-set is dose, which is a continuous variable. For the purpose of this example we'll create a categorical variable called Range where studies are classified based on Dose. The groups are Low (29 to 45), Moderate (48 to 64), and High (66 to 82).

The variable looks like this

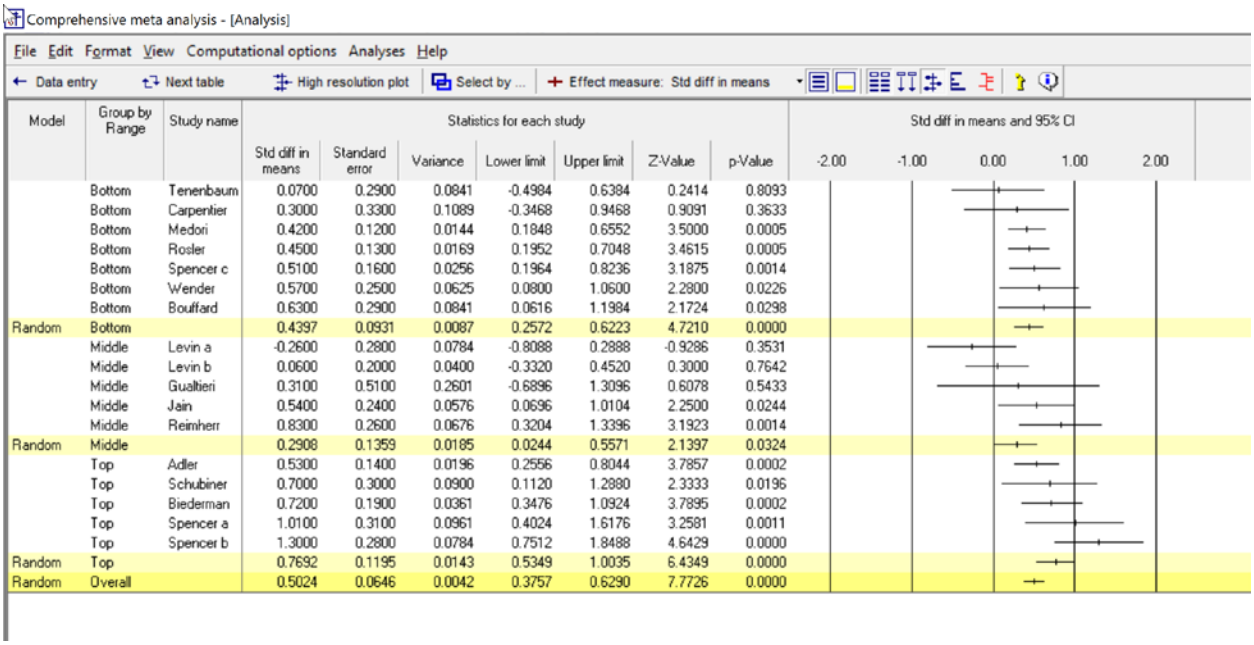


	Study name	Std diff in means	Standard error	Group-A N (Optional)	Group-B N (Optional)	Effect direction	Std diff in means	Std Err	Variance	Dose	Range
1	Spencer c	0.510	0.160			Auto	0.510	0.160	0.026	29.8	Bottom
2	Rosler	0.450	0.130			Auto	0.450	0.130	0.017	41.2	Bottom
3	Medori	0.420	0.120						0.014	42.0	Bottom
4	Wender	0.570	0.250						0.063	43.2	Bottom
5	Bouffard	0.630	0.290						0.084	45.0	Bottom
6	Tenenbaum	0.070	0.290						0.084	45.0	Bottom
7	Carpentier	0.300	0.330						0.109	45.0	Bottom
8	Gualtieri	0.310	0.510						0.260	48.7	Middle
9	Levin b	0.060	0.200						0.040	50.0	Middle
10	Jain	0.540	0.240						0.058	56.8	Middle
11	Levin a	-0.260	0.280						0.078	60.0	Middle
12	Reimherr	0.830	0.260						0.068	64.0	Middle
13	Spencer a	1.010	0.310						0.096	66.5	Top
14	Adler	0.530	0.140						0.020	67.7	Top
15	Schubiner	0.700	0.300						0.090	78.8	Top
16	Biederman	0.720	0.190						0.036	80.9	Top
17	Spencer b	1.300	0.280						0.078	82.0	Top
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											

Note that Range must be defined as a Moderator variable > Categorical

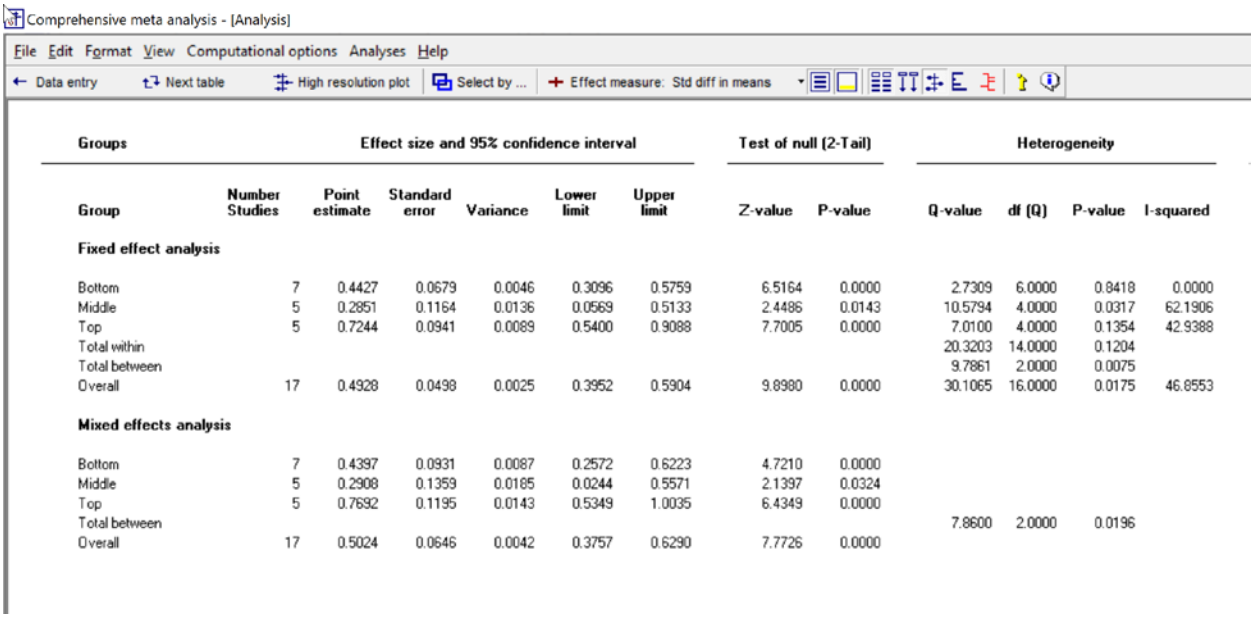
We could perform a subgroups analysis using Range as the moderator.

The analysis would yield the following results



Model	Group by Range	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Bottom	Tenenbaum	0.0700	0.2900	0.0841	-0.4984	0.6384	0.2414	0.8093
	Bottom	Carpentier	0.3000	0.3300	0.1089	-0.3468	0.9468	0.9091	0.3633
	Bottom	Medori	0.4200	0.1200	0.0144	0.1848	0.6552	3.5000	0.0005
	Bottom	Rosler	0.4500	0.1300	0.0169	0.1952	0.7048	3.4615	0.0005
	Bottom	Spencer c	0.5100	0.1600	0.0256	0.1964	0.8236	3.1875	0.0014
	Bottom	Wender	0.5700	0.2500	0.0625	0.0800	1.0600	2.2800	0.0226
	Bottom	Bouffard	0.6300	0.2900	0.0841	0.0616	1.1984	2.1724	0.0298
Random	Bottom		0.4397	0.0931	0.0087	0.2572	0.6223	4.7210	0.0000
	Middle	Levin a	-0.2600	0.2800	0.0784	-0.8088	0.2888	-0.9286	0.3531
	Middle	Levin b	0.0600	0.2000	0.0400	-0.3320	0.4520	0.3000	0.7642
	Middle	Gualtieri	0.3100	0.5100	0.2601	-0.6896	1.3096	0.6078	0.5433
	Middle	Jain	0.5400	0.2400	0.0576	0.0696	1.0104	2.2500	0.0244
	Middle	Reimherr	0.8300	0.2600	0.0676	0.3204	1.3396	3.1923	0.0014
Random	Middle		0.2908	0.1359	0.0185	0.0244	0.5571	2.1397	0.0324
	Top	Adler	0.5300	0.1400	0.0196	0.2556	0.8044	3.7857	0.0002
	Top	Schubiner	0.7000	0.3000	0.0900	0.1120	1.2880	2.3333	0.0196
	Top	Biederman	0.7200	0.1900	0.0361	0.3476	1.0924	3.7895	0.0002
	Top	Spencer a	1.0100	0.3100	0.0961	0.4024	1.6176	3.2581	0.0011
	Top	Spencer b	1.3000	0.2800	0.0784	0.7512	1.8488	4.6429	0.0000
Random	Top		0.7692	0.1195	0.0143	0.5349	1.0035	6.4349	0.0000
Random	Overall		0.5024	0.0646	0.0042	0.3757	0.6290	7.7726	0.0000

Figure 69



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												
Bottom	7	0.4427	0.0679	0.0046	0.3096	0.5759	6.5164	0.0000	2.7309	6.0000	0.8418	0.0000
Middle	5	0.2851	0.1164	0.0136	0.0569	0.5133	2.4486	0.0143	10.5794	4.0000	0.0317	62.1906
Top	5	0.7244	0.0941	0.0089	0.5400	0.9088	7.7005	0.0000	7.0100	4.0000	0.1354	42.9388
Total within									20.3203	14.0000	0.1204	
Total between									9.7861	2.0000	0.0075	
Overall	17	0.4928	0.0498	0.0025	0.3952	0.5904	9.8980	0.0000	30.1065	16.0000	0.0175	46.8553
Mixed effects analysis												
Bottom	7	0.4397	0.0931	0.0087	0.2572	0.6223	4.7210	0.0000				
Middle	5	0.2908	0.1359	0.0185	0.0244	0.5571	2.1397	0.0324				
Top	5	0.7692	0.1195	0.0143	0.5349	1.0035	6.4349	0.0000				
Total between									7.8600	2.0000	0.0196	
Overall	17	0.5024	0.0646	0.0042	0.3757	0.6290	7.7726	0.0000				

Figure 70

The row for Bottom shows the mean effect size for this subgroup is 0.4397. The Z-value of 4.7210 and p-value of < 0.0001 test the null hypothesis that the mean effect size in this group is 0.0000. We reject the null hypothesis and conclude that the effect size in this group is greater than zero.

The row for Middle shows the mean effect size for this subgroup is 0.2908. The Z-value of 2.1397 and p-value of 0.0324 test the null hypothesis that the mean effect size in this group is 0.0000. We reject the null hypothesis and conclude that the effect size in this group is greater than zero.

The row for Top shows the mean effect size for this subgroup is 0.7692. The Z-value of 6.4349 and p-value of < 0.0001 test the null hypothesis that the mean effect size in this group is 0.0000. We reject the null hypothesis and conclude that the effect size in this group is greater than zero.

The row labelled "Total between" displays a test of the null hypothesis that the true mean effect size in all three groups is the same. The Q-value is 7.8600 with 2 degrees of freedom and a corresponding p-value of 0.0196. We reject the null hypothesis and conclude that the effect size varies among subgroups.

If we wanted to compare the mean effect size for the three subgroups we could stop here. However, if we wanted to compare these effect sizes while adjusting for other covariates (for example, to see if this pattern remains when we control for SUD) we would need to use a regression. Given the modest number of studies, this is intended only as an exercise.

Dummy variables

CMA is able to create the dummy variables for *Range* automatically.

In the regression module (Figure 71),

- Click on Show Covariates [A]
- Click on Range [B]
- Click on Edit reference group [C]
- Select [Bottom] [D]
- Click [Add to main screen] [E]

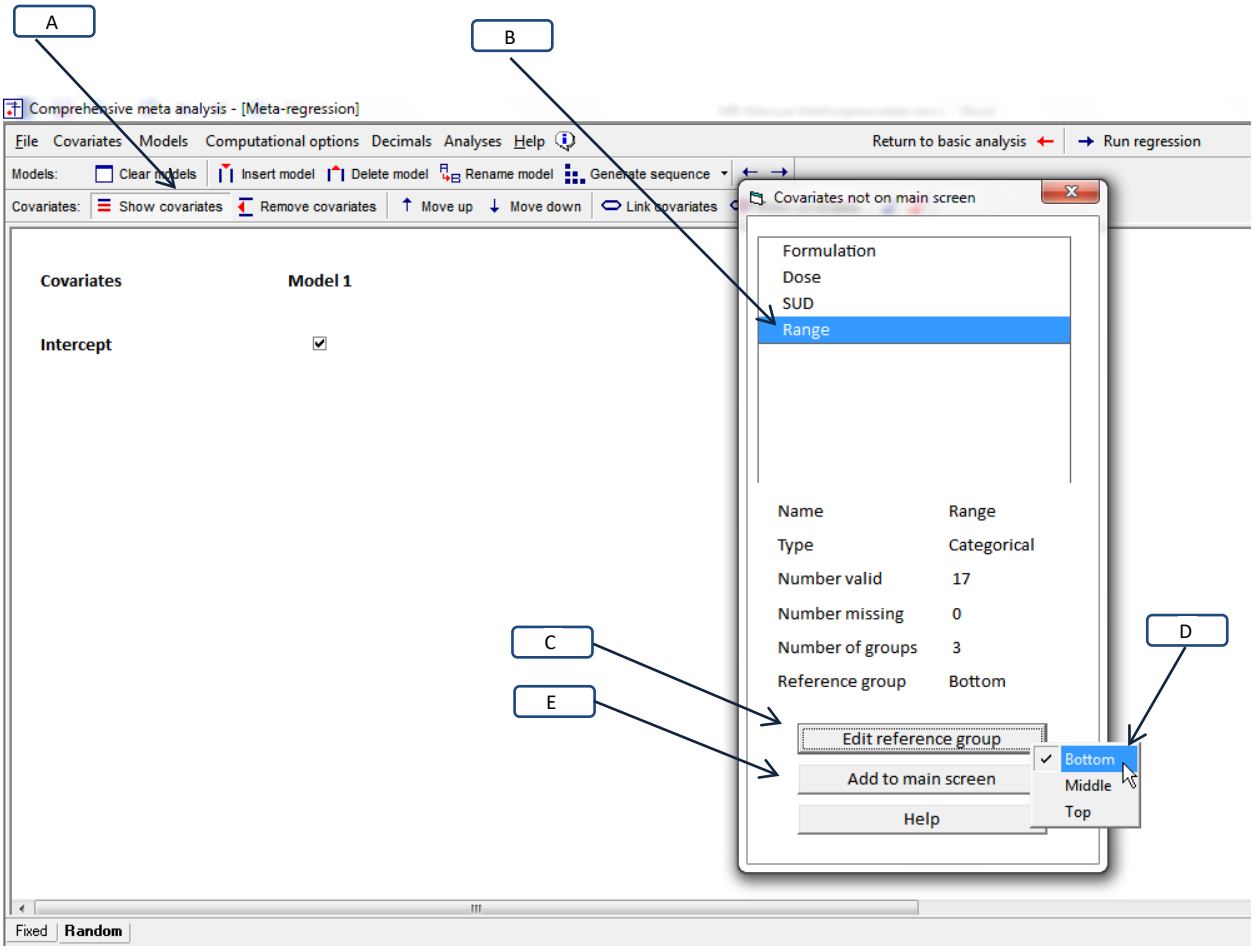


Figure 71 | Creating dummy variables

Since we've set "Bottom" as the reference group, the Dummy variables are "Middle" and "Top". In Figure 72 [F] the program creates these and adds them to the variable list.

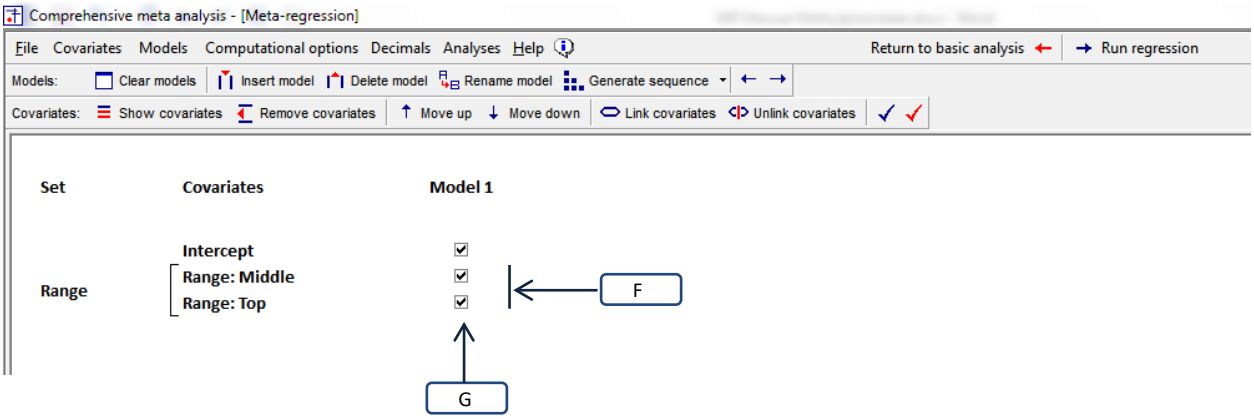


Figure 72 | Creating dummy variables

As always, tick the boxes [G] to include these variables in the current predictive model. Tick *either* of the two boxes, and the other will be ticked automatically. This is because the two represent Range, and not because they belong to the same set.

The two dummy variables are Moderate and High. Following the conventions proposed by Cohen, studies are coded "1" if they belong to the dummy-variable's group name. Thus,

- A study is coded 1 for "Middle" if the range is "Middle", or 0 otherwise.
- A study is coded 1 for "Top" if the range is "Top", or 0 otherwise.

Working with the "Set"

When the program creates a series of dummy variables to represent a categorical covariate, it automatically defines these dummy variables as a "Set". In Figure 72, and Figure 73 the program has added a column labeled "Set". In this column we see the label "Range", which refers to the categorical variable. Brackets indicate the two dummy variables that represent Range.

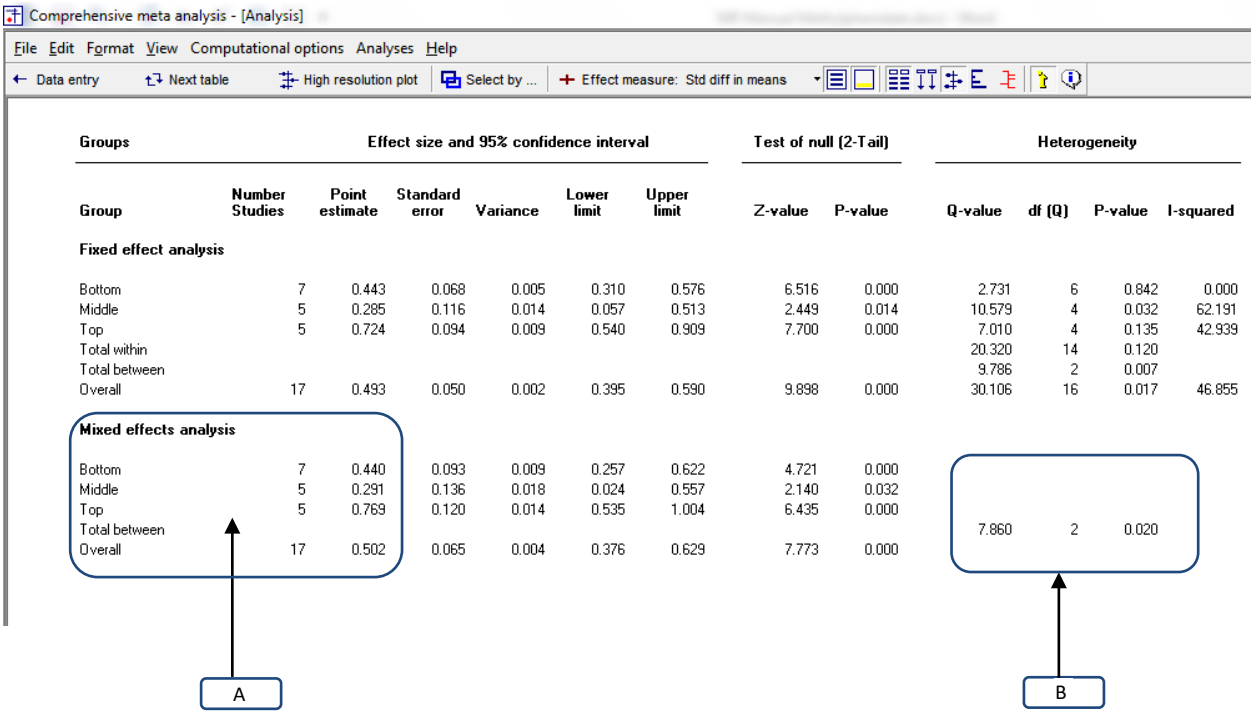
In our example, Range (in the form of dummy variables) is the only covariate in the equation, and so the test of the set is identical to the test of the model. The test of the set [D] yields a *Q*-value of 7.8600 with 2 *df* and $p = 0.0196$. Equivalently, the test of the model [B] yields a *Q*-value of 7.8600 with 2 *df* and $p = 0.0196$. Therefore, in this example we really didn't need to present statistics for the set. We could have simply relied on the statistics for the model.

However, this choice only exists when the variables in the set are the *only* variables in the model. By contrast, when there are additional covariates in the model, the test of the set is quite different from the test of the model. For example, suppose the model includes Range (dummy coded) and also Formulation. The Set would test the *unique* impact of Range, with Formulation partialled. By contrast, the model would test the *combined* impact of Range and Formulation. These are two entirely different issues.

BEFORE "SET" SHOW CORRESPONDENCE W/ SUBGROUPS

p-value for each (and model) in subgroup vs regression

ADD SECTION ON HOW TO SET SUBGROUP OPTIONS



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												
Bottom	7	0.443	0.068	0.005	0.310	0.576	6.516	0.000	2.731	6	0.842	0.000
Middle	5	0.285	0.116	0.014	0.057	0.513	2.449	0.014	10.579	4	0.032	62.191
Top	5	0.724	0.094	0.009	0.540	0.909	7.700	0.000	7.010	4	0.135	42.939
Total within									20.320	14	0.120	
Total between									9.786	2	0.007	
Overall	17	0.493	0.050	0.002	0.395	0.590	9.898	0.000	30.106	16	0.017	46.855
Mixed effects analysis												
Bottom	7	0.440	0.093	0.009	0.257	0.622	4.721	0.000				
Middle	5	0.291	0.136	0.018	0.024	0.557	2.140	0.032				
Top	5	0.769	0.120	0.014	0.535	1.004	6.435	0.000				
Total between									7.860	2	0.020	
Overall	17	0.502	0.065	0.004	0.376	0.629	7.773	0.000				

Figure 73 | Subgroups | Dose range

- The mean effect size for each subgroup [A] is the same as the numbers from the regression
- The *Q*-value for the model [B] is the same as the numbers from the regression

In this example, Range (or rather the dummy variables that represent Range) is the only covariate in the regression. Therefore, the intercept is simply the mean effect for the reference group and the coefficients represent the difference in mean effects. If the regression model included other covariates then all the statistics would be adjusted for the other covariates.

Note. For the subgroups analysis, Computational options > Random and mixed-effect options must be set to pool estimates of T^2.

AND FIXED AT BOTTOM

Using sets with Dummy variables

One common use of sets is the case where we use two or more dummy variables to represent a categorical covariate. In this case the program will create the set automatically.

For example, the data set includes a categorical covariate called “Range” which classifies studies based on the dosage range as “Bottom”, “Middle”, or “Top” range of dosages. If we enter Range into the

model the program automatically creates two dummy variables. In this example the Reference group is Bottom, and the dummy variables are called Middle and Top.

At this point the program automatically defined these two dummy variables as a set, as shown in Figure 74 [A].

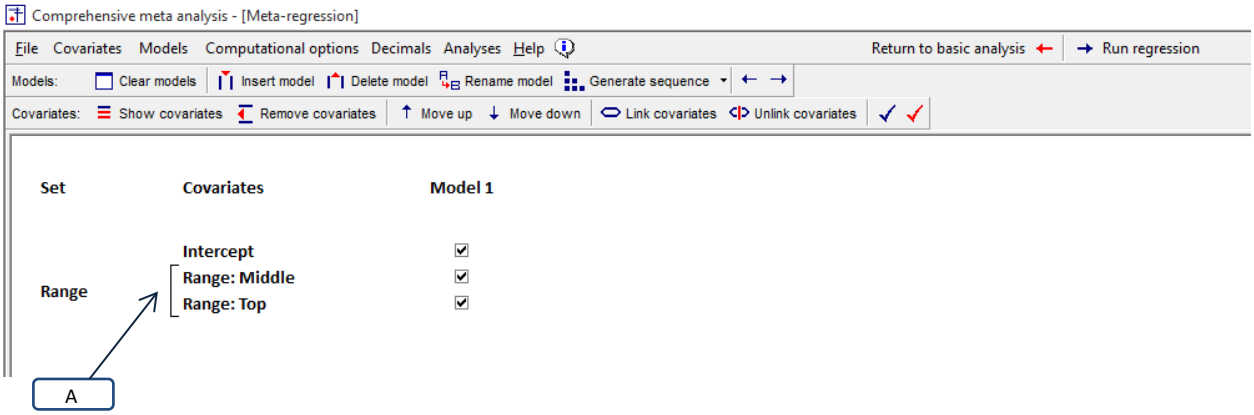


Figure 74

The results are shown in Figure 75.

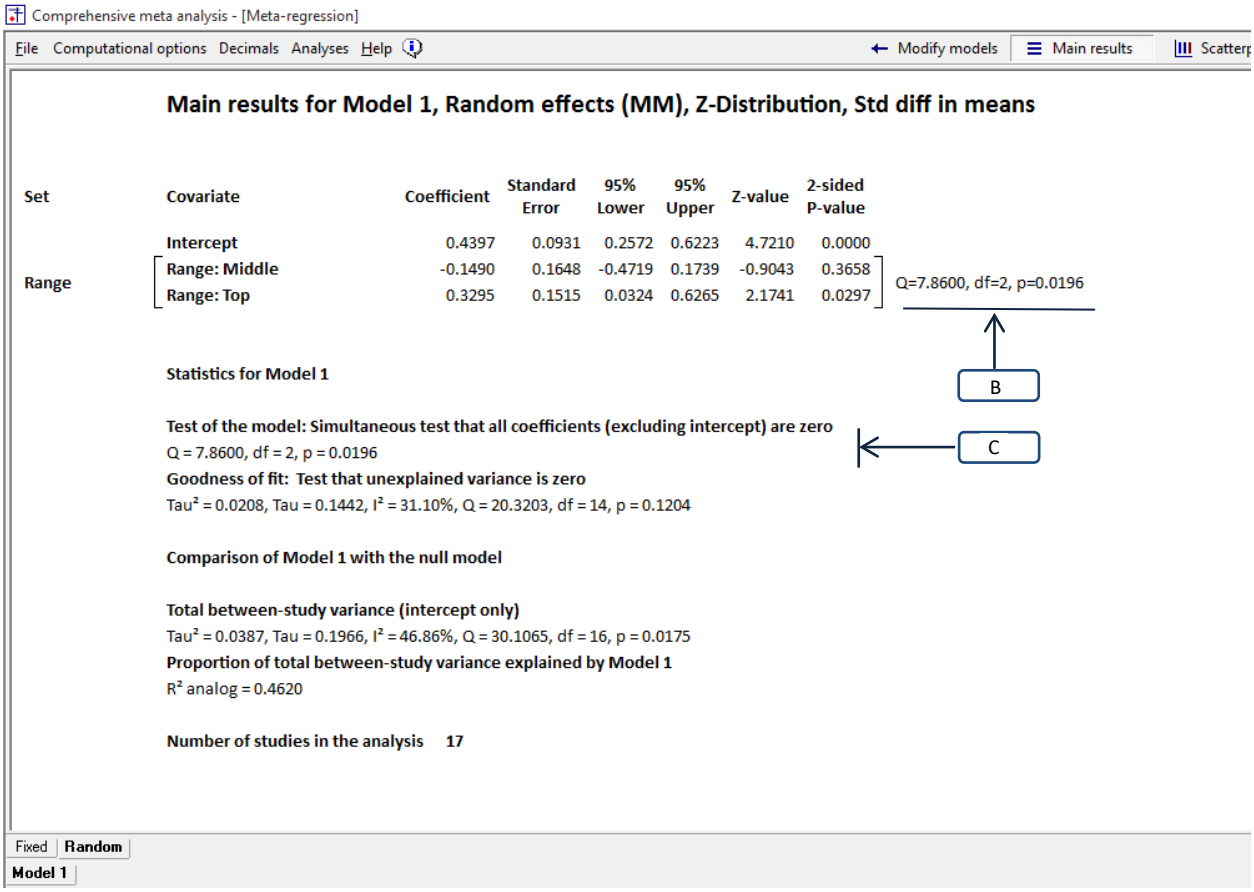


Figure 75 | Regression | Main results | Assessing the impact of a set

If we want to ask specifically if the Middle range differs from the others, or if the Top range differs from the others, then we would look to the corresponding row. That is,

- The statistics for Range: Middle tells us if employing a Middle-range dosage is related to effect size (when all other covariates are partialled). The coefficient is −0.1490, the Z-value is −0.9043 and the p-value is 0.3658.

- The statistics for Range: Top tells us if the use of a Top-range dosage is related to effect size (when all other covariates are partialled). The coefficient is 0.3295, the Z-value is 2.1741 and the p-value is 0.0297.

More often, we are not concerned with a specific category, but rather intend to ask if Range in general (Bottom vs. Middle vs. Top) is related to the effect size. In this case we would look at the Set. That is,

- The statistics for the set [B] tell us if Range as a whole (that is, the use of Bottom, Middle, or Top range) is related to effect size. Here, Q=7.8600, df=2, p=0.0196.

When the covariates in the set are the only covariates in the prediction equation (as they are in this example) the statistics for the set will be identical to the statistics for the model. Specifically, the *Q*-value for the set [B] and the *Q*-value for the model [C] are both 7.8600 with *df* = 2 and *p* = 0.0196. Therefore, in this example (Figure 75) we could have simply employed the test of the model as the test of Allocation.

However, that is not the case when the model includes additional covariates.

Consider Figure 76, where the model includes SUD as well as the two dummy-variables for Range.

- The statistics for the set [D] address the impact of Range with SUD partialled. For this, *Q* = 6.7524, df=2, p=0.0342.

- By contrast, the statistics for the model [E] address the joint impact of Range *and* SUD. For this, *Q* = 14.0441, df=3, p=0.0028.

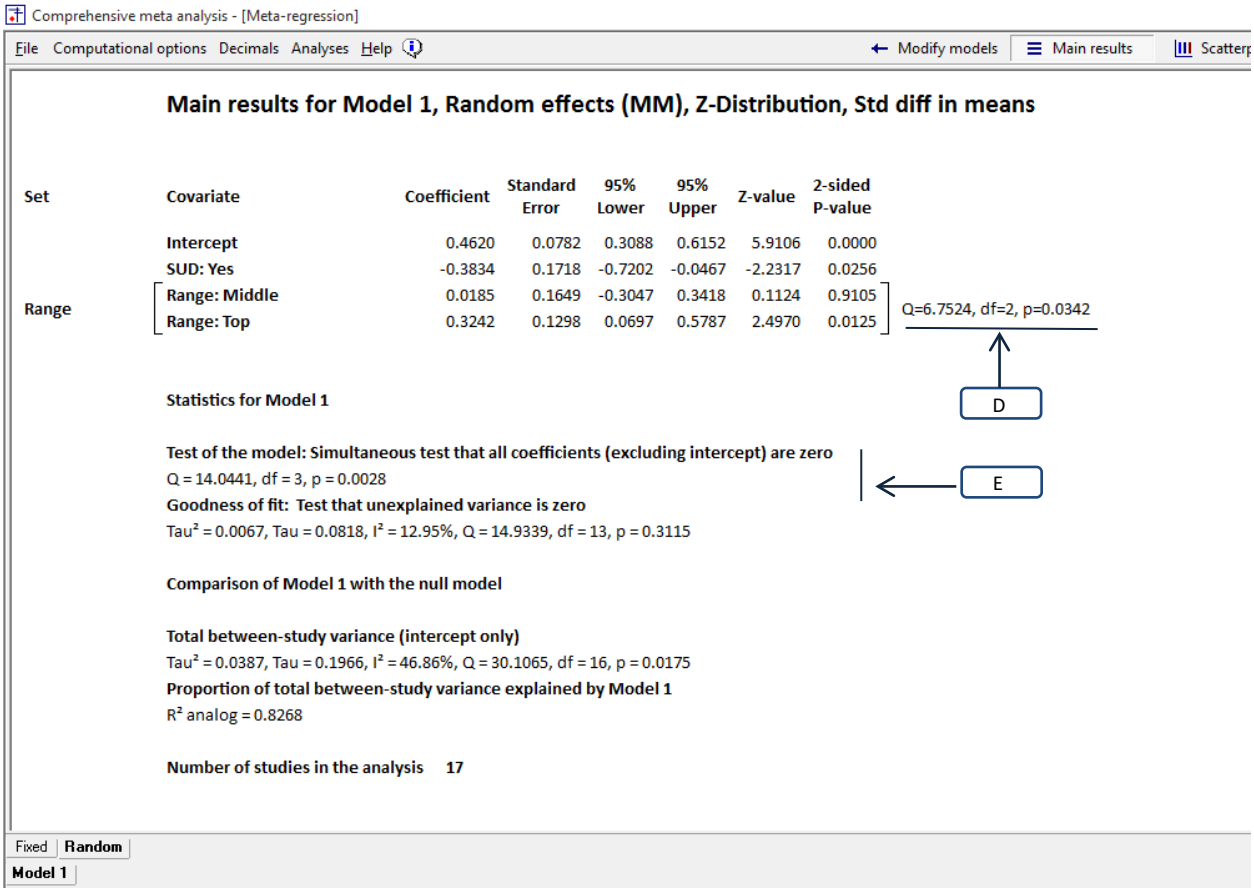


Figure 76 | Main results | Assessing the impact of a set

Note that this is intended as a theoretical exercise only. In order for the analysis to be useful we would need to have more studies. Also, we would need to have some studies with and without SUD at each level of Range. Since there are only 4 studies with SUD patients, it's clear that we don't have that.

CREATING DUMMY VARIABLES MANUALLY

Above, we showed how to create dummy variables automatically. While the option to create dummy variables automatically works well in most cases, there are several cases where you'll need to create the dummy-variables manually. The following are some examples where you would want to create dummy-variables manually.

Interactions

Suppose you have a categorical variable that is represented by a dummy variable A, and you want to assess the impact of that variable and also its interaction with another variable B. You'll need to work with the variables A, B, and AB. In this case it will be easier to create A (and then AB) manually.

Alternate coding schemes

Dummy-coding is only one of the options possible for categorical covariates. Texts on multiple regression discuss other options, such as effects-coding and contrast-coding. You can use any of these coding schemes, but you'll need to create the dummy variables manually.

Regressions with no intercept

The automatic coding scheme is only available when you include the intercept in the equation. When you omit the intercept (see next chapter) the coding scheme changes (in that case, for *m* groups you need *m* rather than *m* – 1 dummy variables) and the automatic function is not available.

If you create the dummy variables manually, you'll also need to define these as a set manually. This is discussed in Motivating example SUD

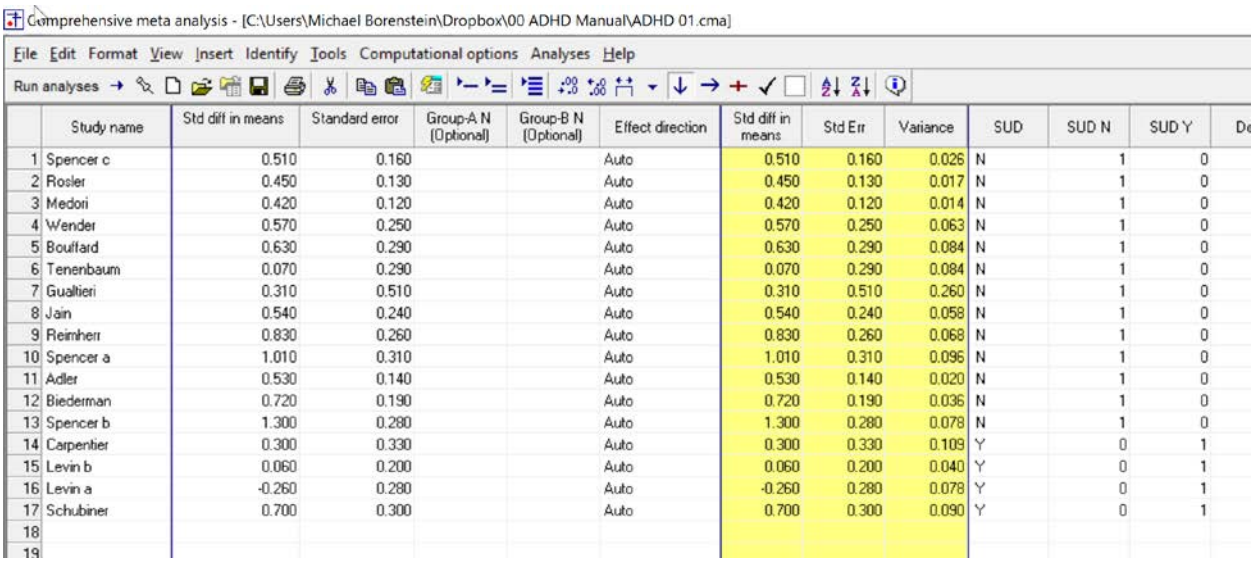
The data set includes a covariate called SUD, which stands for substance abuse disorder. This variable is defined as being categorical, and each study is coded as either No (excluded SUD patients) or Yes (Included SUD patients).

Since SUD is categorical, it cannot be inserted directly into the analysis. Rather, we need to create a numerical covariate corresponding to Formulation and use this in the analysis. For purposes of this example we will create two variables, called SUD-No and SUD-Yes.

Whereas SUD is defined as a Categorical moderator, the new variables must be defined as Integer moderators since they will be entered directly into the regression.

For SUD-No, studies are coded 1 (No) or 0 (Others)
For SUD-Yes, studies are coded 1 (Yes) or 0 (Others)

In this example we sorted the studies by SUD, then entered 1 and 0 for the first 13 studies, followed by 0 and 1 for the last 4 studies.



	Study name	Std diff in means	Standard error	Group-A N (Optional)	Group-B N (Optional)	Effect direction	Std diff in means	Std Err	Variance	SUD	SUD N	SUD Y
1	Spencer c	0.510	0.160			Auto	0.510	0.160	0.026	N	1	0
2	Rosler	0.450	0.130			Auto	0.450	0.130	0.017	N	1	0
3	Medori	0.420	0.120			Auto	0.420	0.120	0.014	N	1	0
4	Wender	0.570	0.250			Auto	0.570	0.250	0.063	N	1	0
5	Boulfard	0.630	0.290			Auto	0.630	0.290	0.084	N	1	0
6	Tenenbaum	0.070	0.290			Auto	0.070	0.290	0.084	N	1	0
7	Gualtieri	0.310	0.510			Auto	0.310	0.510	0.260	N	1	0
8	Jain	0.540	0.240			Auto	0.540	0.240	0.058	N	1	0
9	Reimherr	0.830	0.260			Auto	0.830	0.260	0.068	N	1	0
10	Spencer a	1.010	0.310			Auto	1.010	0.310	0.096	N	1	0
11	Adler	0.530	0.140			Auto	0.530	0.140	0.020	N	1	0
12	Biederman	0.720	0.190			Auto	0.720	0.190	0.036	N	1	0
13	Spencer b	1.300	0.280			Auto	1.300	0.280	0.078	N	1	0
14	Carpentier	0.300	0.330			Auto	0.300	0.330	0.109	Y	0	1
15	Levin b	0.060	0.200			Auto	0.060	0.200	0.040	Y	0	1
16	Levin a	-0.260	0.280			Auto	-0.260	0.280	0.078	Y	0	1
17	Schubiner	0.700	0.300			Auto	0.700	0.300	0.090	Y	0	1
18												
19												

Figure 77 | Data-entry | Dummy variables for Continuous and Intermittent

There is a basic difference in how we use dummy variables when we include vs. omit the intercept. Consider a variable with m categories. In this example studies are coded as either SUD-No or SUD-Yes ($m = 2$).

When we include the intercept, we include $m - 1$ dummy variables. We would include either SUD-No or SUD-Yes, but not both

When we omit the intercept, we include m dummy variables. We would include both SUD-No and SUD-Yes.

While the program is able to create dummy-variables automatically for the case where we include the intercept, this function is not available for the case where we omit the intercept. In this example we'll be using the dummy-variables that we created manually.

When we set up the regression we omit the intercept [C] and include both SUD N and SUD Y as covariates [D].

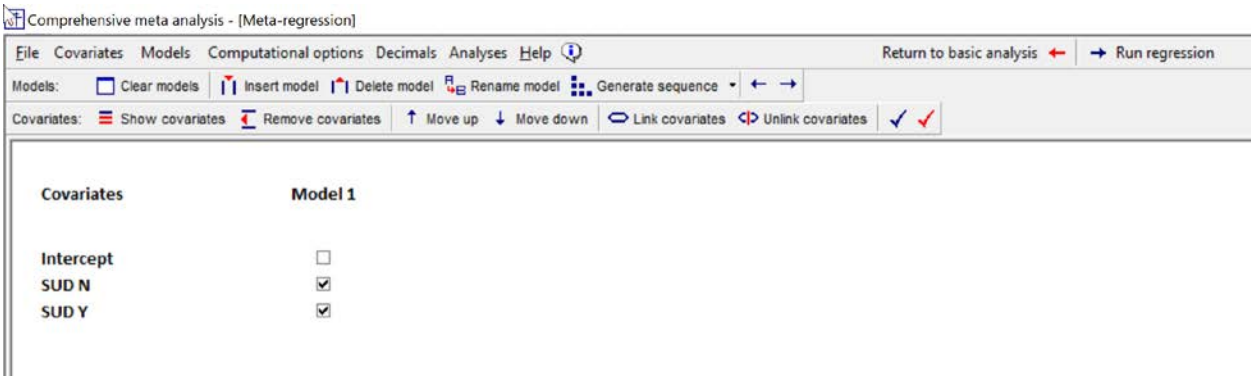


Figure 78 | Regression | Setup | No intercept

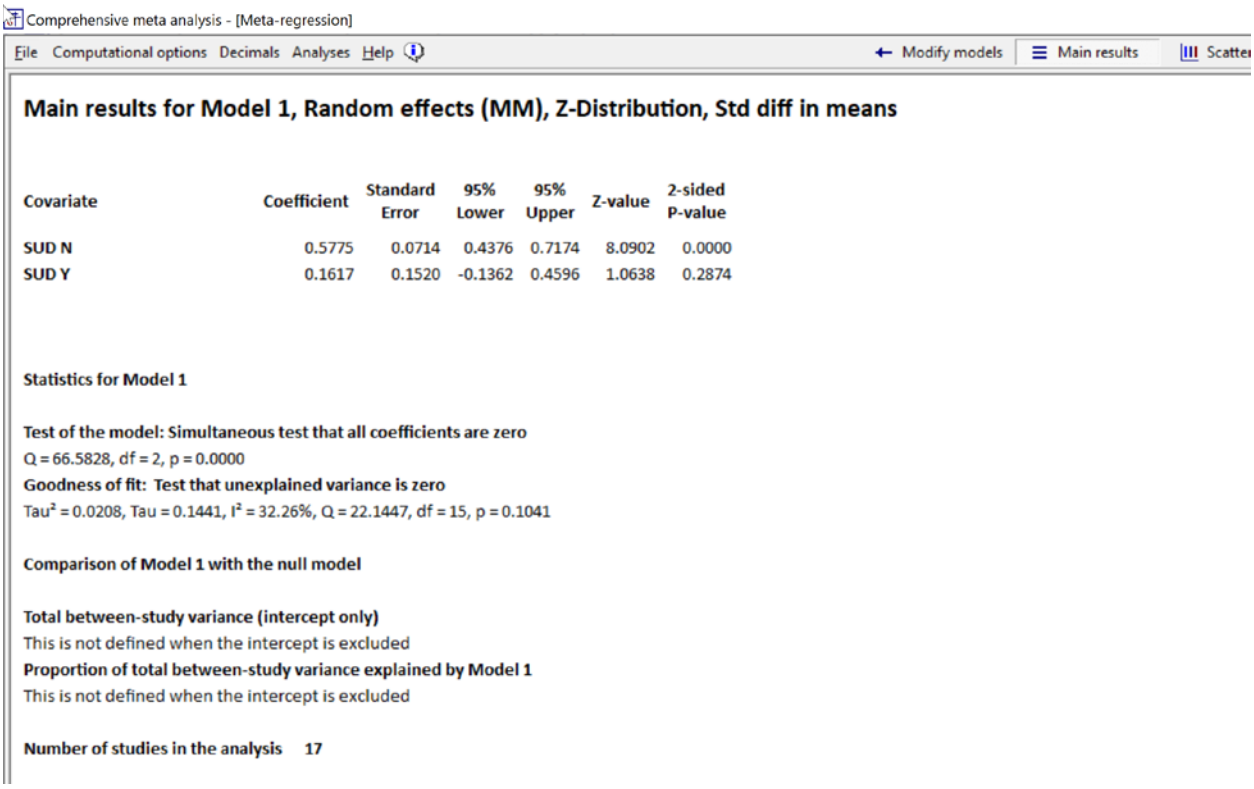


Figure 79 | Regression | Main results | No intercept

Because we have omitted the intercept, the statistics reflect the mean effect size for each group.

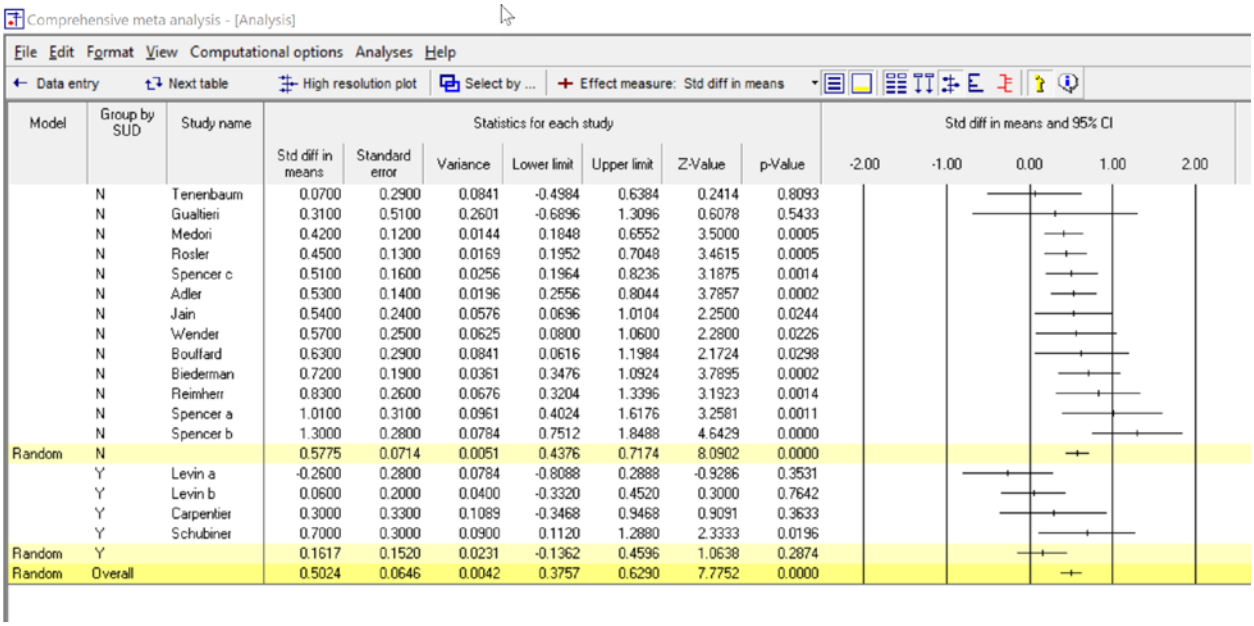
For the SUD N subgroup [AA] the mean effect size is 0.5775 with SE = 0.0714. The test addresses the question "Is this effect size zero?" and yields $Z = 8.0902$, $p < 0.0001$.

For the SUD Y subgroup [BB] the mean effect size is 0.1617 with SE = 0.1520. The test addresses the question "Is this effect size zero?" and yields $Z = 1.0638$, $p = 0.2874$.

For context, compare these results of this analysis to the results we would see if we ran a subgroups analysis.

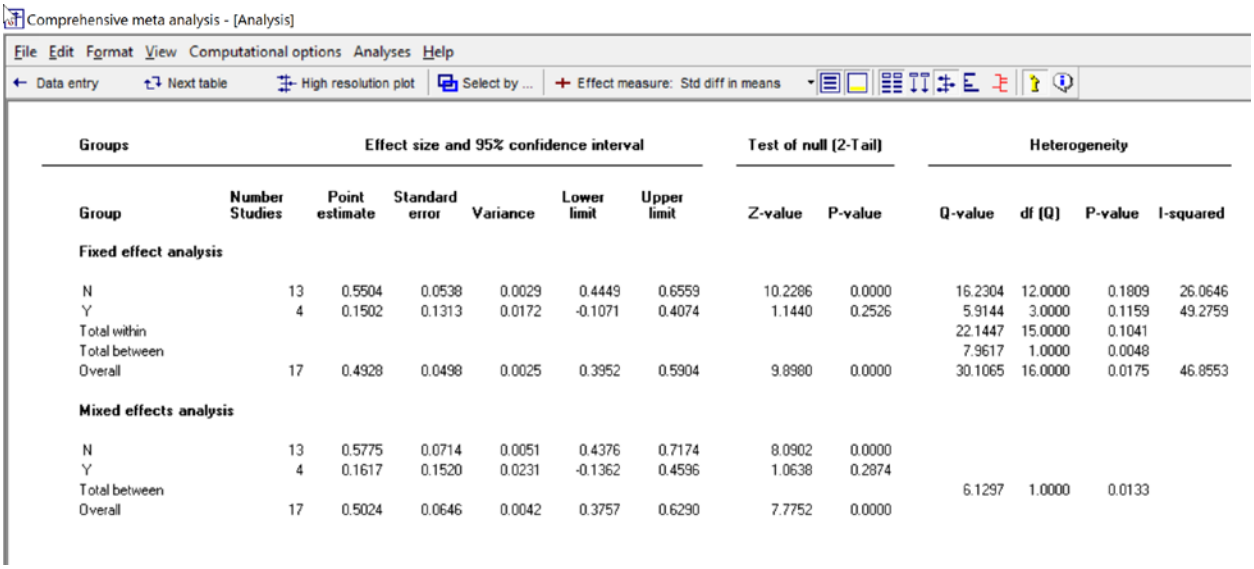
On the main analysis screen click Computational options > Mixed and random effects options and then select the option to "pool within-group estimates of tau-squared". At the bottom of this box, selected Fixed (even though we are using the random-effects model). See appendix

Click Computational options > Group by and group by Formulation
Select the "Random" tab at the bottom of the screen.



Model	Group by SUD	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	N	Tenenbaum	0.0700	0.2900	0.0841	-0.4984	0.6384	0.2414	0.8093
	N	Gualtieri	0.3100	0.5100	0.2601	-0.6896	1.3096	0.6078	0.5433
	N	Medori	0.4200	0.1200	0.0144	0.1848	0.6552	3.5000	0.0005
	N	Rosler	0.4500	0.1300	0.0169	0.1952	0.7048	3.4615	0.0005
	N	Spencer c	0.5100	0.1600	0.0256	0.1964	0.8236	3.1875	0.0014
	N	Adler	0.5300	0.1400	0.0196	0.2556	0.8044	3.7857	0.0002
	N	Jain	0.5400	0.2400	0.0576	0.0696	1.0104	2.2500	0.0244
	N	Wender	0.5700	0.2500	0.0625	0.0800	1.0600	2.2800	0.0226
	N	Boulfard	0.6300	0.2900	0.0841	0.0616	1.1984	2.1724	0.0298
	N	Biederman	0.7200	0.1900	0.0361	0.3476	1.0924	3.7895	0.0002
	N	Reimherr	0.8300	0.2600	0.0676	0.3204	1.3396	3.1923	0.0014
	N	Spencer a	1.0100	0.3100	0.0961	0.4024	1.6176	3.2581	0.0011
	N	Spencer b	1.3000	0.2800	0.0784	0.7512	1.8488	4.6429	0.0000
Random	N		0.5775	0.0714	0.0051	0.4376	0.7174	8.0902	0.0000
	Y	Levin a	-0.2600	0.2800	0.0784	-0.8088	0.2888	-0.9286	0.3531
	Y	Levin b	0.0600	0.2000	0.0400	-0.3320	0.4520	0.3000	0.7642
	Y	Carpentier	0.3000	0.3300	0.1089	-0.3468	0.9468	0.9091	0.3633
	Y	Schubiner	0.7000	0.3000	0.0900	0.1120	1.2880	2.3333	0.0196
Random	Y		0.1617	0.1520	0.0231	-0.1362	0.4596	1.0638	0.2874
Random	Overall		0.5024	0.0646	0.0042	0.3757	0.6290	7.7752	0.0000

Figure 80 | Subgroups | Continuous vs. Intermittent



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												
N	13	0.5504	0.0538	0.0029	0.4449	0.6559	10.2286	0.0000	16.2304	12.0000	0.1809	26.0646
Y	4	0.1502	0.1313	0.0172	-0.1071	0.4074	1.1440	0.2526	5.9144	3.0000	0.1159	49.2759
Total within									22.1447	15.0000	0.1041	
Total between									7.9617	1.0000	0.0048	
Overall	17	0.4928	0.0498	0.0025	0.3952	0.5904	9.8980	0.0000	30.1065	16.0000	0.0175	46.8553
Mixed effects analysis												
N	13	0.5775	0.0714	0.0051	0.4376	0.7174	8.0902	0.0000				
Y	4	0.1617	0.1520	0.0231	-0.1362	0.4596	1.0638	0.2874				
Total between									6.1297	1.0000	0.0133	
Overall	17	0.5024	0.0646	0.0042	0.3757	0.6290	7.7752	0.0000				

Figure 81 | Subgroups | Continuous vs. Intermittent

Figure 66 and Figure 67 show the results of this analysis

For the SUD N subgroup [AA] the mean effect size is 0.5775 with SE = 0.0714. The test addresses the question “Is this effect size zero?” and yields $Z = 8.0902$, $p < 0.0001$.

For the SUD Y subgroup [BB] the mean effect size is 0.1617 with SE = 0.1520. The test addresses the question “Is this effect size zero?” and yields $Z = 1.0638$, $p = 0.2874$.

These statistics are identical to the ones reported for the regression.

In fact, in this example we could indeed have used either subgroups or regression and obtained the same results. However, regression offers options that subgroups analysis does not. Specifically, regression allows us to include additional covariates. If we do so, the regression will give us mean effect for each category controlling for other covariates, as well as the standard error of the adjusted effect.

Notes.

The test for an overall effect is not the same in the two analyses. In the regression the test of the model is a test of the null hypothesis that the mean for all groups is zero. The *Q*-value of 66.5828 with 2 degrees of freedom yields a *p*-value of < 0.0001. We reject the null and conclude that the mean effect size is probably not zero in at least one of the groups. In the subgroups analysis the test of the overall effect is a test of the null that the (weighted) mean effect size across subgroups is zero. The Z-value is 7.7752 with a corresponding p-value of < 0.0001.

In this chapter we addressed the case where we have one categorical variable (and any number of continuous variables). The process of creating dummy variables becomes more complicated when we have two or more categorical variables. This is beyond the scope of this manual.

Working with Sets of covariates.

To create dummy-variables manually, on the data-entry sheet create a column for a moderator and then (critically) define the moderator type as integer or decimal. Then, enter a value for each study.

Step-by-step examples are given in the appendix

Note.

If you create Dummy variables manually, these must be classified as Moderator > Integer or as Moderator > Decimal. Do not classify them as Moderator > Categorical.

WHEN DOES IT MAKE SENSE TO OMIT THE INTERCEPT

In any regression we have the option to either include or omit the intercept from the prediction equation. The decision to omit the intercept usually arises when we are working with a categorical variable, and we will discuss the issue in this context. The decision to include or omit the intercept fundamentally affects the issues addressed by the analysis.

When we *include* the intercept in a regression with categorical covariates

- Coefficients reflect *differences* in effect size across categories.
- Tests of a covariate address the question "Is the covariate related to effect size?"
- The model tests the null hypothesis that *no* covariate is related to effect size.

When we *omit* the intercept in a regression with categorical covariates

- Coefficients reflect *the absolute effect sizes* within categories.
- Tests of a covariate address the question "Is the effect size zero?" in this category.
- The model tests the null hypothesis that all groups have a mean of zero.

So, when we're interested in the impact of covariates on the effect size we'll include the intercept. On the other hand, if we want to estimate the effect size for specific subgroups (rather than estimate the difference between them) we would omit the intercept. This is the case, for example, if we want to estimate these effects so we can use them in a network meta-analysis.

In this chapter we cover two issues.

- First, we will address a technical issue about coding categorical variables.
- Second, we will show how to interpret an analysis where the intercept is omitted.

As always, we caution the reader that this chapter is not a comprehensive treatment of the topic. We assume that the reader is familiar with these issues from multiple regression in primary studies. Our goal here is to review the key concepts, and show how they can be applied in meta-analysis.

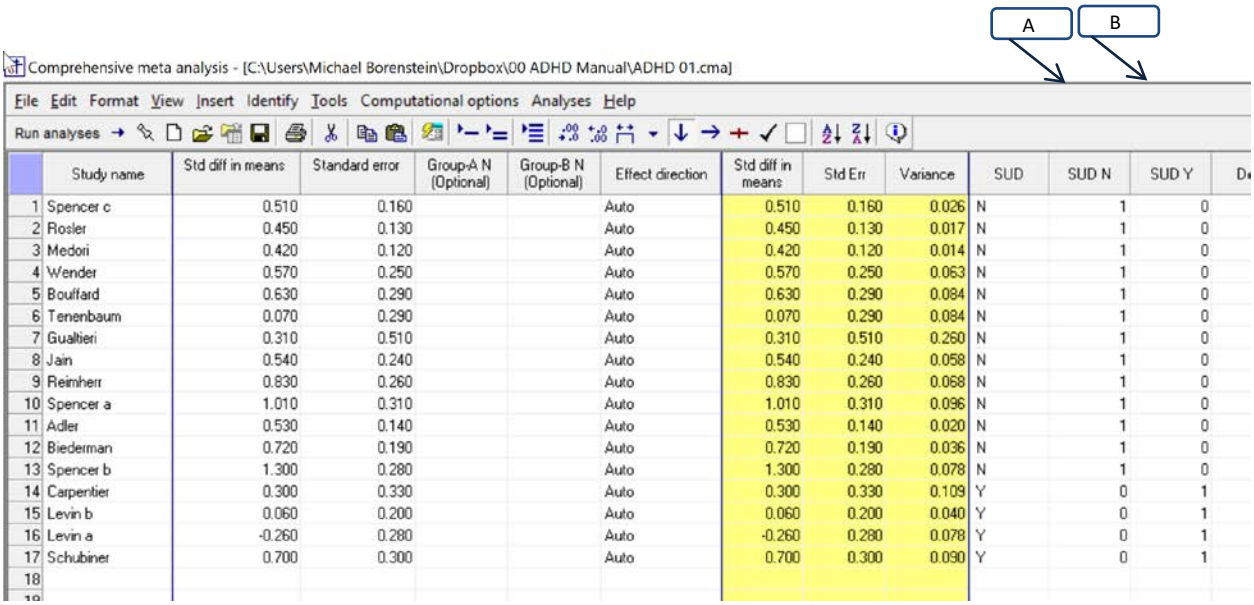
Motivating example SUD

The data set includes a covariate called *SUD,* which stands for *substance abuse disorder*. This variable is defined as being categorical, and each study is coded as either No (excluded SUD patients) or Yes (Included SUD patients).

Since SUD is categorical, it cannot be inserted directly into the analysis. Rather, we need to create a numerical covariate corresponding to Formulation and use this in the analysis. For purposes of this example we will create two variables, called SUD-No and SUD-Yes.

Whereas SUD is defined as a Categorical moderator, the new variables must be defined as Integer moderators since they will be entered directly into the regression.

For SUD-No, studies are coded 1 (No) or 0 (Others)
For SUD-Yes, studies are coded 1 (Yes) or 0 (Others)



	Study name	Std diff in means	Standard error	Group-A N (Optional)	Group-B N (Optional)	Effect direction	Std diff in means	Std Err	Variance	SUD	SUD N	SUD Y
1	Spencer c	0.510	0.160			Auto	0.510	0.160	0.026	N	1	0
2	Rosler	0.450	0.130			Auto	0.450	0.130	0.017	N	1	0
3	Medori	0.420	0.120			Auto	0.420	0.120	0.014	N	1	0
4	Wender	0.570	0.250			Auto	0.570	0.250	0.063	N	1	0
5	Boulfard	0.630	0.290			Auto	0.630	0.290	0.084	N	1	0
6	Tenenbaum	0.070	0.290			Auto	0.070	0.290	0.084	N	1	0
7	Gualtieri	0.310	0.510			Auto	0.310	0.510	0.260	N	1	0
8	Jain	0.540	0.240			Auto	0.540	0.240	0.058	N	1	0
9	Reimherr	0.830	0.260			Auto	0.830	0.260	0.068	N	1	0
10	Spencer a	1.010	0.310			Auto	1.010	0.310	0.096	N	1	0
11	Adler	0.530	0.140			Auto	0.530	0.140	0.020	N	1	0
12	Biederman	0.720	0.190			Auto	0.720	0.190	0.036	N	1	0
13	Spencer b	1.300	0.280			Auto	1.300	0.280	0.078	N	1	0
14	Carpentier	0.300	0.330			Auto	0.300	0.330	0.109	Y	0	1
15	Levin b	0.060	0.200			Auto	0.060	0.200	0.040	Y	0	1
16	Levin a	-0.260	0.280			Auto	-0.260	0.280	0.078	Y	0	1
17	Schubiner	0.700	0.300			Auto	0.700	0.300	0.090	Y	0	1
18												

Figure 82 | Data-entry | Dummy variables for Continuous and Intermittent

There is a basic difference in how we use dummy variables when we include vs. omit the intercept. Consider a variable with *m* categories. In this example studies are coded as either SUD-No or SUD-Yes (*m* = 2).

- When we *include* the intercept, we include *m*-1 dummy variables. We would include *either* SUD-No *or* SUD-Yes, but not both

- When we *omit* the intercept, we include *m* dummy variables. We would include *both* SUD-No *and* SUD-Yes.

- While the program is able to create dummy-variables automatically for the case where we include the intercept, this function is not available for the case where we omit the intercept. In this example we'll be using the dummy-variables that we created manually.

When we set up the regression we omit the intercept [C] and include *both* SUD N and SUD Y as covariates [D].

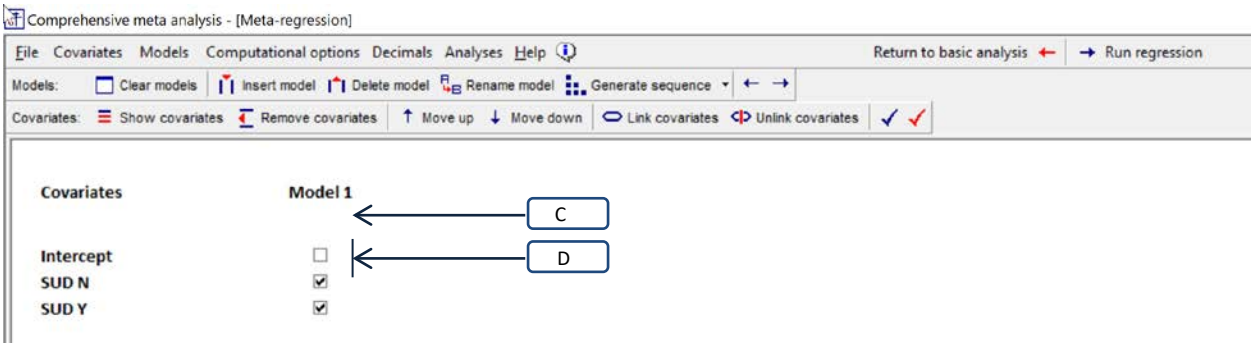


Figure 83 | Regression | Setup | No intercept

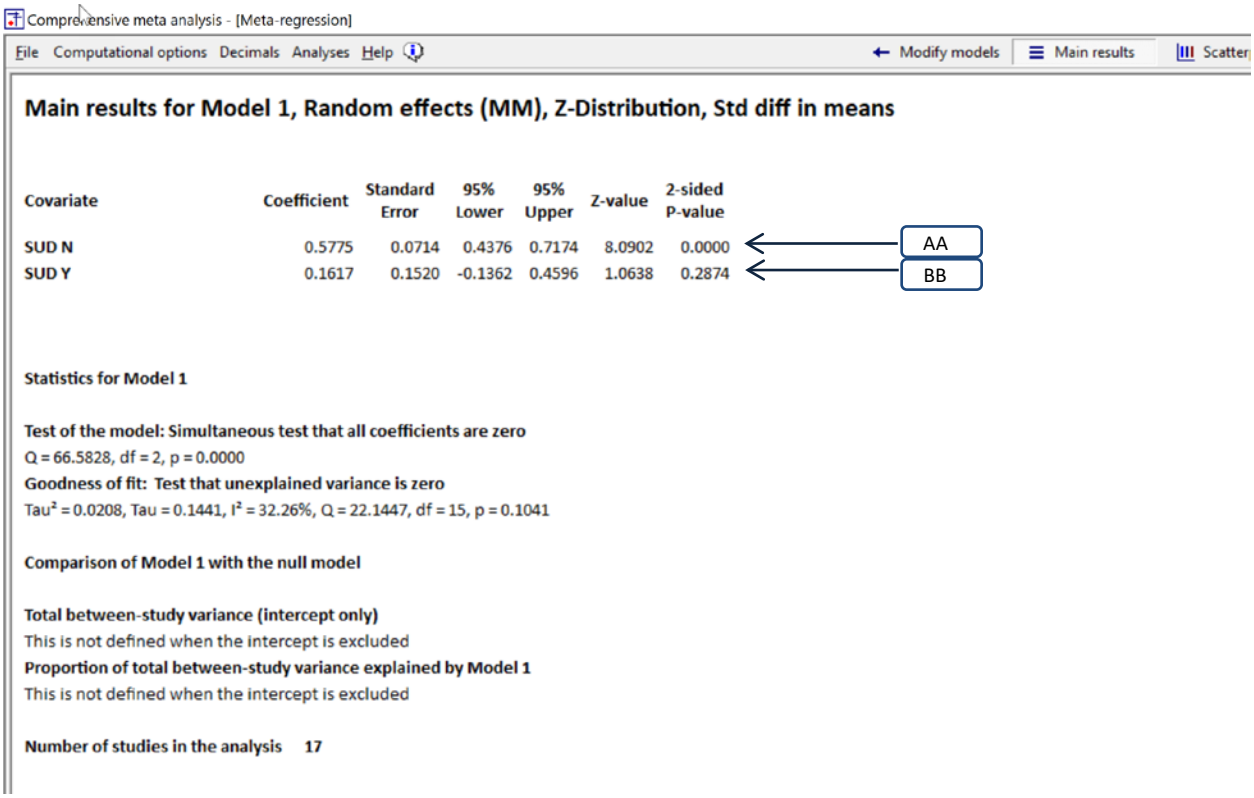


Figure 84 | Regression | Main results | No intercept

Because we have omitted the intercept, the statistics reflect the mean effect size for each group.

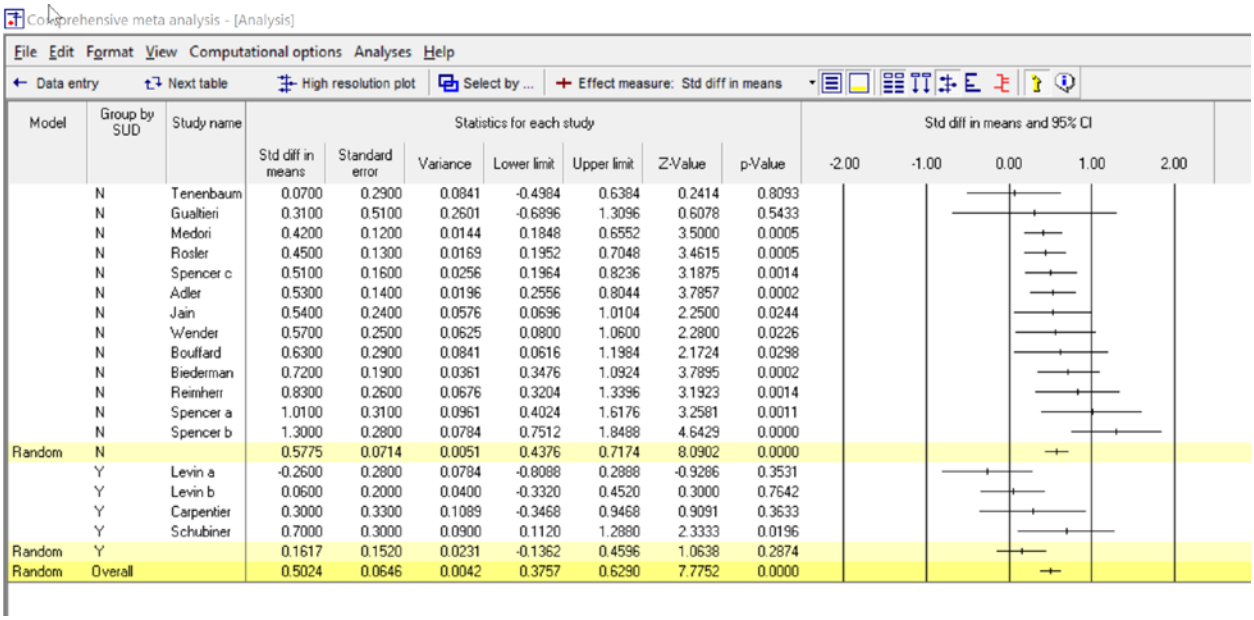
- For the SUD N subgroup [AA] the mean effect size is 0.5775 with *SE* = 0.0714 The test addresses the question "Is this effect size zero?" and yields *Z* = 8.0902, *p* < 0.0001.

- For the SUD Y subgroup [BB] the mean effect size is 0.1617 with *SE* = 0.1520. The test addresses the question "Is this effect size zero?" and yields *Z* = 1.0638, *p* = 0.2874.

For context, compare these results of this analysis to the results we would see if we ran a subgroups analysis.

On the main analysis screen click Computational options > Mixed and random effects options and then select the option to "pool within-group estimates of tau-squared" (see appendix).

- Click Computational options > Group by and group by Formulation
- Select the "Random" tab at the bottom of the screen.



Model	Group by SUD	Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	N	Tenenbaum	0.0700	0.2900	0.0841	-0.4984	0.6384	0.2414	0.8093
	N	Gualtieri	0.3100	0.5100	0.2601	-0.6896	1.3096	0.6078	0.5433
	N	Medori	0.4200	0.1200	0.0144	0.1848	0.6552	3.5000	0.0005
	N	Rosler	0.4500	0.1300	0.0169	0.1952	0.7048	3.4615	0.0005
	N	Spencer c	0.5100	0.1600	0.0256	0.1964	0.8236	3.1875	0.0014
	N	Adler	0.5300	0.1400	0.0196	0.2556	0.8044	3.7857	0.0002
	N	Jain	0.5400	0.2400	0.0576	0.0696	1.0104	2.2500	0.0244
	N	Wender	0.5700	0.2500	0.0625	0.0800	1.0600	2.2800	0.0226
	N	Bouffard	0.6300	0.2900	0.0841	0.0616	1.1984	2.1724	0.0298
	N	Biederman	0.7200	0.1900	0.0361	0.3476	1.0924	3.7895	0.0002
	N	Reimherr	0.8300	0.2600	0.0676	0.3204	1.3396	3.1923	0.0014
	N	Spencer a	1.0100	0.3100	0.0961	0.4024	1.6176	3.2581	0.0011
	N	Spencer b	1.3000	0.2800	0.0784	0.7512	1.8488	4.6429	0.0000
Random	N		0.5775	0.0714	0.0051	0.4376	0.7174	8.0902	0.0000
	Y	Levin a	-0.2600	0.2800	0.0784	-0.8088	0.2888	-0.9286	0.3531
	Y	Levin b	0.0600	0.2000	0.0400	-0.3320	0.4520	0.3000	0.7642
	Y	Carpentier	0.3000	0.3300	0.1089	-0.3468	0.9468	0.9091	0.3633
	Y	Schubiner	0.7000	0.3000	0.0900	0.1120	1.2880	2.3333	0.0196
Random	Y		0.1617	0.1520	0.0231	-0.1362	0.4596	1.0638	0.2874
Random	Overall		0.5024	0.0646	0.0042	0.3757	0.6290	7.7752	0.0000

Figure 85 | Subgroups | Continuous vs. Intermittent

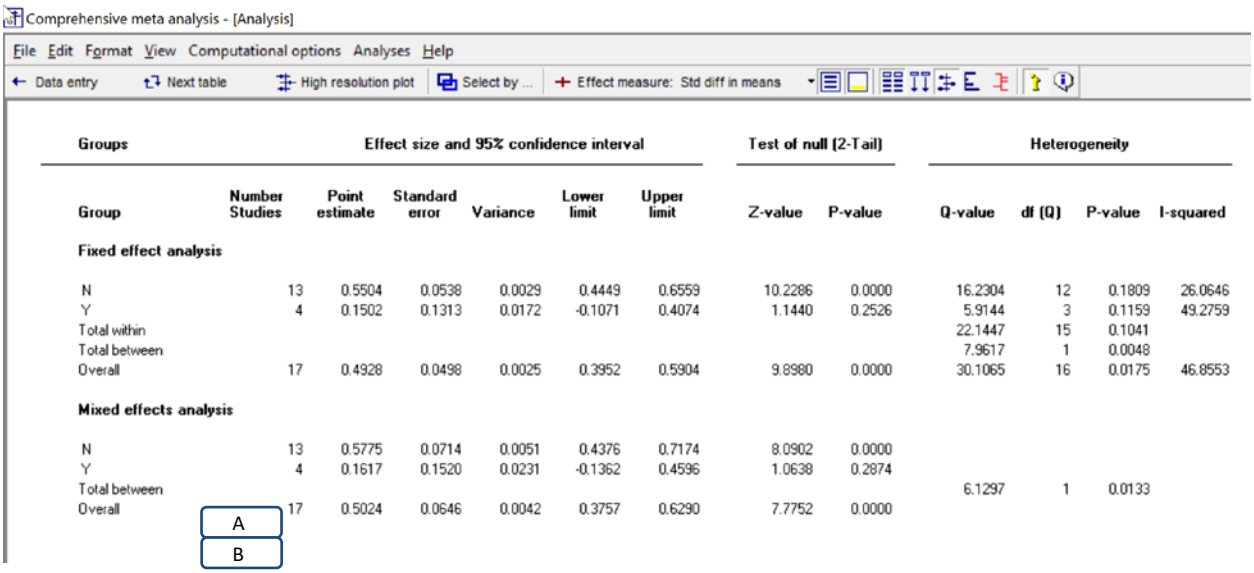


Figure 86 | Subgroups | Continuous vs. Intermittent

Error! Reference source not found. and **Error! Reference source not found.** show the results of this analysis

- For the SUD N subgroup [AA] the mean effect size is 0.5775 with *SE* = 0.0714 The test addresses the question "Is this effect size zero?" and yields $Z = 8.0902$, $p < 0.0001$.

- For the SUD Y subgroup [BB] the mean effect size is 0.1617 with *SE* = 0.1520. The test addresses the question "Is this effect size zero?" and yields $Z = 1.0638$, $p = 0.2874$.

These statistics are identical to the ones reported for the regression.

In fact, in this example we could indeed have used either subgroups or regression and obtained the same results. However, regression offers options that subgroups analysis does not. Specifically, regression allows us to include additional covariates. If we do so, the regression will give us mean effect for each category controlling for other covariates, as well as the standard error of the adjusted effect.

Notes.

The test for an overall effect is *not* the same in the two analyses. In the regression the test of the model is a test of the null hypothesis that the mean for *all* groups is zero. The *Q*-value of 66.5828 with 2 degrees of freedom yields a *p*-value of < 0.0001. We reject the null and conclude that the mean effect size is probably not zero in at least one of the groups. In the subgroups analysis the test of the overall effect is a test of the null that the (weighted) mean effect size across subgroups is zero. For this test, the Z-value is 7.7752 and the p-value is < 0.0001.

In this chapter we addressed the case where we have one categorical variable (and any number of continuous variables). The process of creating dummy variables becomes more complicated when we have two or more categorical variables. This is beyond the scope of this manual.

WORKING WITH SETS OF COVARIATES

DEFINING A SET

In regression there are times when we use several covariates to capture a concept. For example

- If we want to assess the relationship between Dose and effect we might include *Dose* and $Dose^2$. The first addresses the linear relationship between dose and effect size, while the latter addresses the curvilinear relationship.
- We would use a set of dummy-variables to represent a categorical variable. For example
- We may have a series of covariates, such as *income* and *education* that (together) represent the impact of socio-economic status.
- We may have a series of covariates such as *dose* and *duration* that (together) represent the intensity of a treatment.
- We may have covariates *A* and *B* and also the interaction between these covariates (*AB*), where the three together represent the impact of this set on outcome.

When we define covariates as a Set, the program reports a test of significance for the Set with all other covariates held constant.

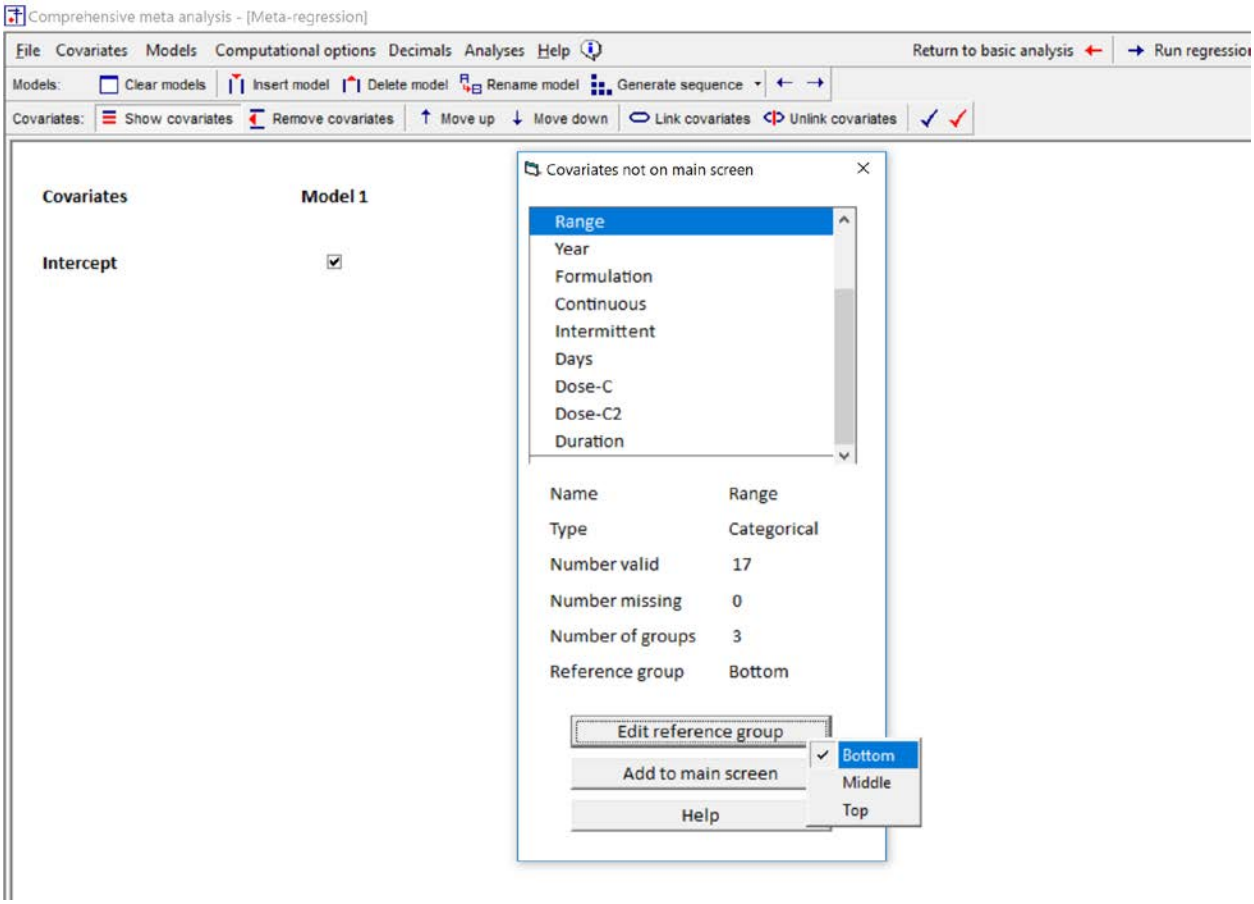
In this chapter we explain how to create a set, and how to interpret the results.

HOW TO CREATE A SET

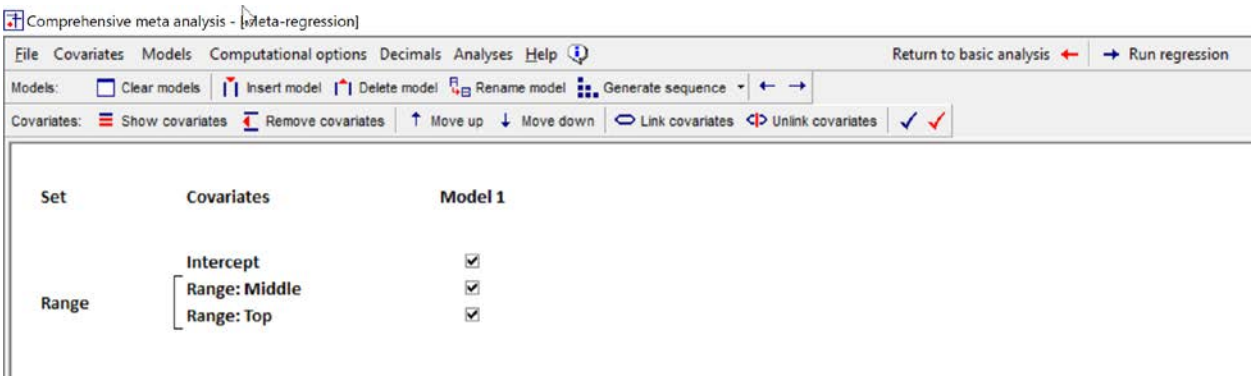
Creating a set for a categorical variable

One common use for sets is the case where we are working with a categorical variable that has three or more categories. If there are m categories, the program will create a set of m-1 dummy variables to represent the variable, and we will want to define these as a set.

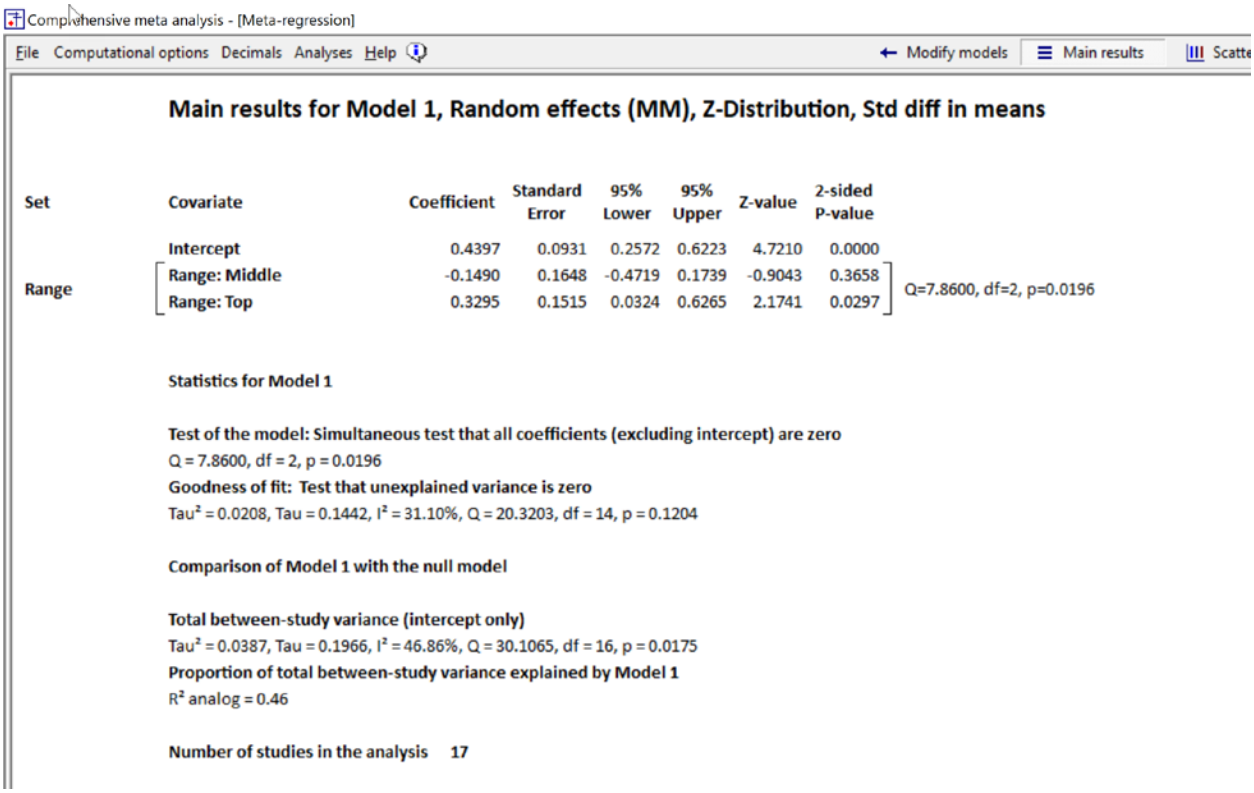
For example, consider the variable called Range. Each study is coded Bottom, Middle, or Top, based on the dosage of medication. (Typically we would use Dose as a continuous variable, and we created these categories for purposes of this exercise).



Click on Range
Edit Reference group, and select Bottom
Add to main screen



The program creates two dummy variables, names them Range: Middle and Range: Top. Additionally, it creates a Set called "Range" and put a bracket around the two dummy variables that make up this set.



The program displays statistics for each covariate.

The coefficient for Middle is -0.1490, which tells us that the mean effect size for studies in the Middle range is 0.1490 points lower than the bottom range (since the bottom range is the reference). The standard error is 0.1648, the confidence interval is -0.4719 to 0.1739, the Z-value is -0.9043 and the p-value is 0.3658. This is not statistically significant, and so we cannot reject the null hypothesis that the mean effect size for the Middle range is the same as the mean effect size for the Bottom range.

The coefficient for Top is 0.3295, which tells us that the mean effect size for studies in the Top range is 0.3295 points higher than the bottom range (since the bottom range is the reference). The standard

error is 0.1515, the confidence interval is 0.0324 to 0.6265, the Z-value is 2.1741 and the p-value is 0.0297. This is statistically significant, and so we can conclude that the mean effect size for the Top range is higher than the mean effect size for the Bottom range.

Finally, we might want to ask about Range, as a whole. We could pose the null hypothesis that the mean effect size is the same for all three levels or (equivalently) that there is no relationship between range and effect size.

This is not addressed by either row alone, but it is addressed by the two rows together, as captured by the set. The Q value for the set is 7.8600 with two degrees of freedom and p=0.0196. We can reject the null hypothesis that the mean effect size is the same for all levels of Range, and conclude that it differs across levels.

While it might seem that only two levels are included in the set, the fact is that all three levels are included. Every study is coded either 1,0 (Middle), 0,1 (Top) or 0,0 (Bottom), so every study's level is identified and included in the analysis.

In this example, since the only covariates are those in the set, the set and the Model are the same. Therefore, the test of the set is identical to the test of the model and the Q-value for the set is the same as the Q value for the test of the model. In this case, therefore, we could have worked with the model. However, the value of the set become apparent if we add another covariate.

For example, we've established that the effect size varies by range, but we think this might be due to a confound with Days. Perhaps the reason that the mean effect size is lower for studies in the middle range is because these tend to be longer than studies in the other ranges. And, if we control for Days, the differences among the ranges would disappear.

To test for this we can add Days to the regression equation as shown here.

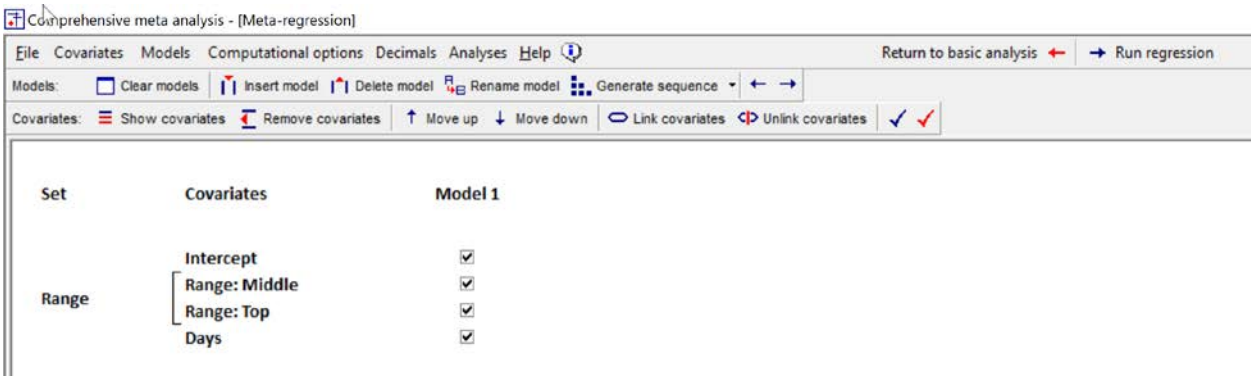
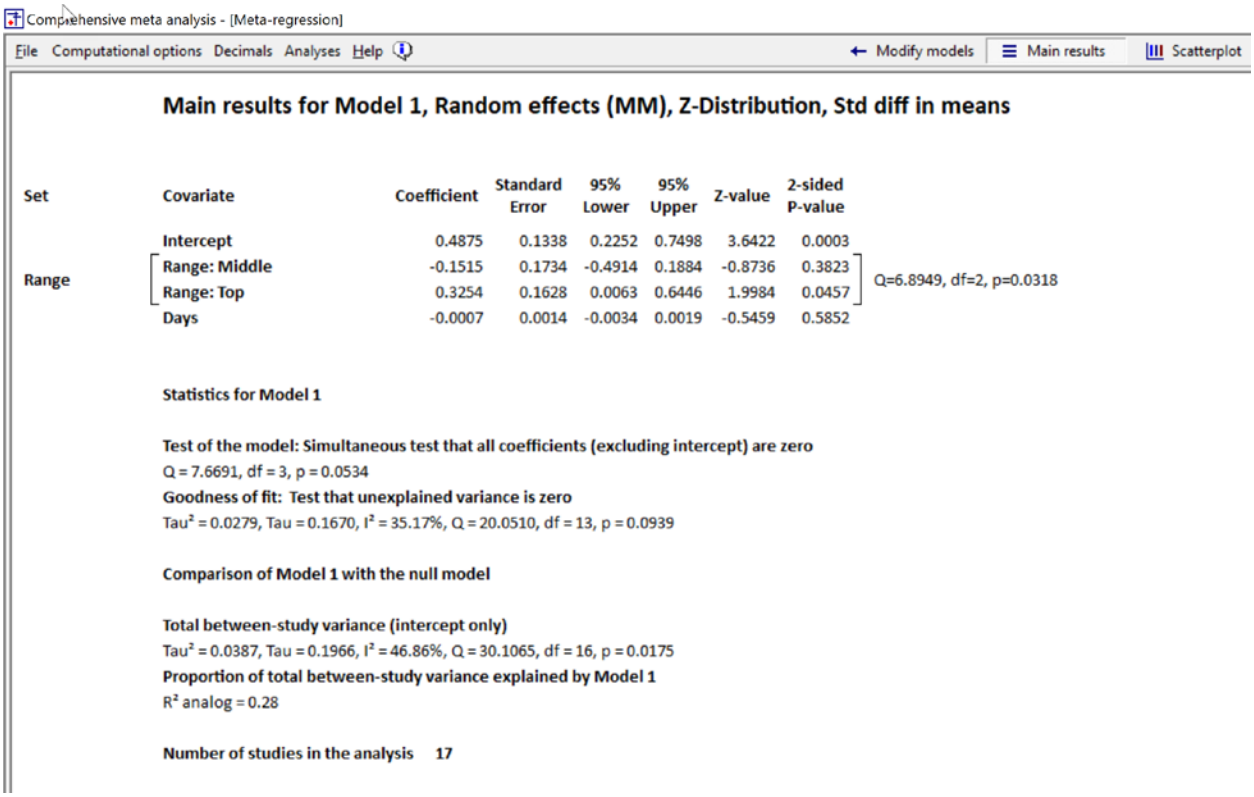


Figure 87



In this case, the statistics for each row and for the set are computed partialing Days. So,

For two studies that lasted the same number of days, the expected mean effect size for studies in the Middle range is 0.1515 points lower than the expected mean for studies in the Bottom range. The standard error is 0.1734 with a confidence interval of -0.4914 to 0.1884, a Z value of -0.8736, and a p-value of 0.3823. This is not statistically significant, so there is no evidence that the mean effect size is different for studies in the Middle range vs. the Bottom range, when Days is held constant.

For two studies that lasted the same number of days, the expected mean effect size for studies in the Top range is 0.3254 points higher than the expected mean for studies in the Bottom range. The standard error is 0.1628 with a confidence interval of 0.0063 to 0.6446, a Z value of 1.9984, and a p-value of 0.0457. This is statistically significant, so we conclude that the mean effect size is higher for studies in the Top range vs. the Bottom range, when Days is held constant.

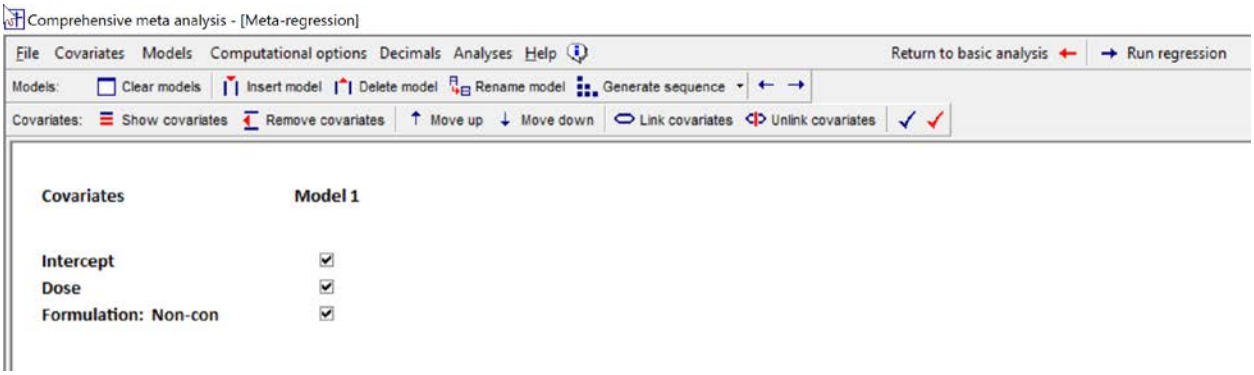
Next, we turn to the set. The Q-value for the set is 6.8949 with two degrees of freedom and p=0.0318. We reject the null hypothesis that the mean effect size is the same for all levels of Range when Days is held constant. We conclude that the treatment is more effective for the studies in some ranges than in others, and that (at least some of) this difference cannot be explained as a confound of Days.

Finally, we can look at the model. The model includes Range and Days. The Q-value for the model is 7.6691 with three degrees of freedom and p=0.0534. If we consider this statistically significant, it tells us that either Range or Days or Both is related to effect size, but does not allow us to isolate the impact of Range. By contract, the Set does allow us to do so.

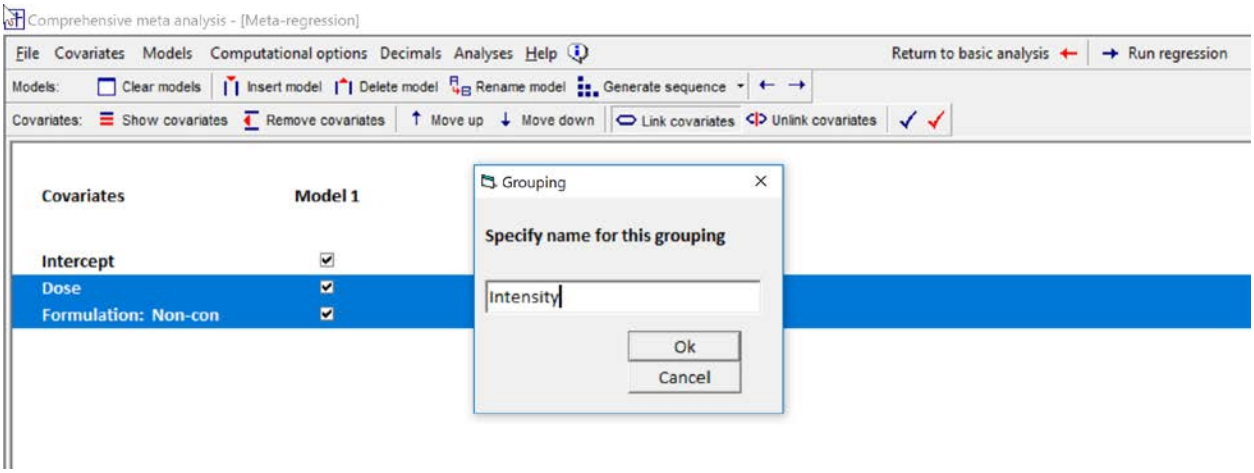
Another case where we might want to create a set is where two variables together may represent the intensity of the intervention. For example, we have a variable called Dose and another called Formulation. Formulation refers to the manner of administration (continuous or non-continuous), and it's possible that non-continuous administration potentiates the impact of the drug. We might want to include the two as a set, to ask if the intensity (defined this way) is related to the effect size.

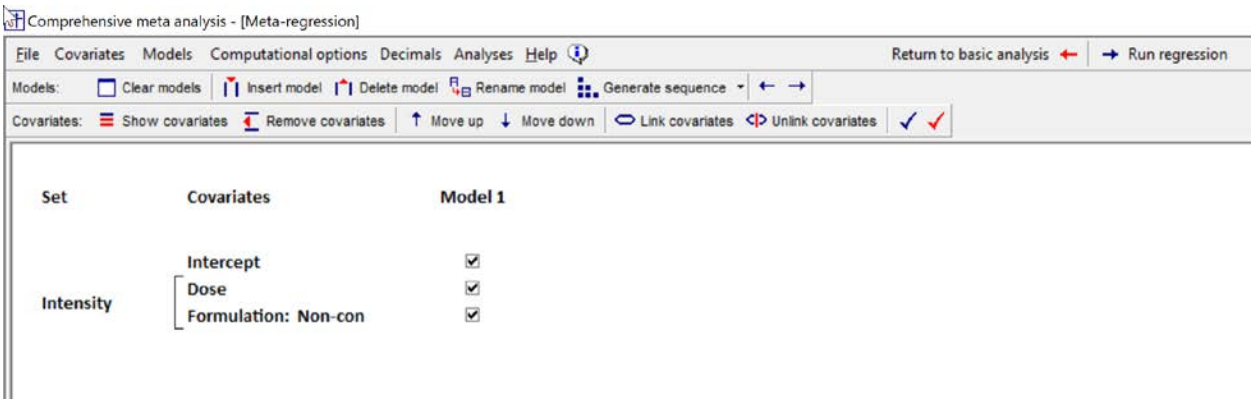
In this case, we will need to create the set manually.

Add Dose to the Model
Add Formulation: Non-con to the Model



Highlight Dose and Formulation: Non-con
Click Link Covariates
Name the grouping Intensity
Click Ok
The program puts a bracket around the covariates





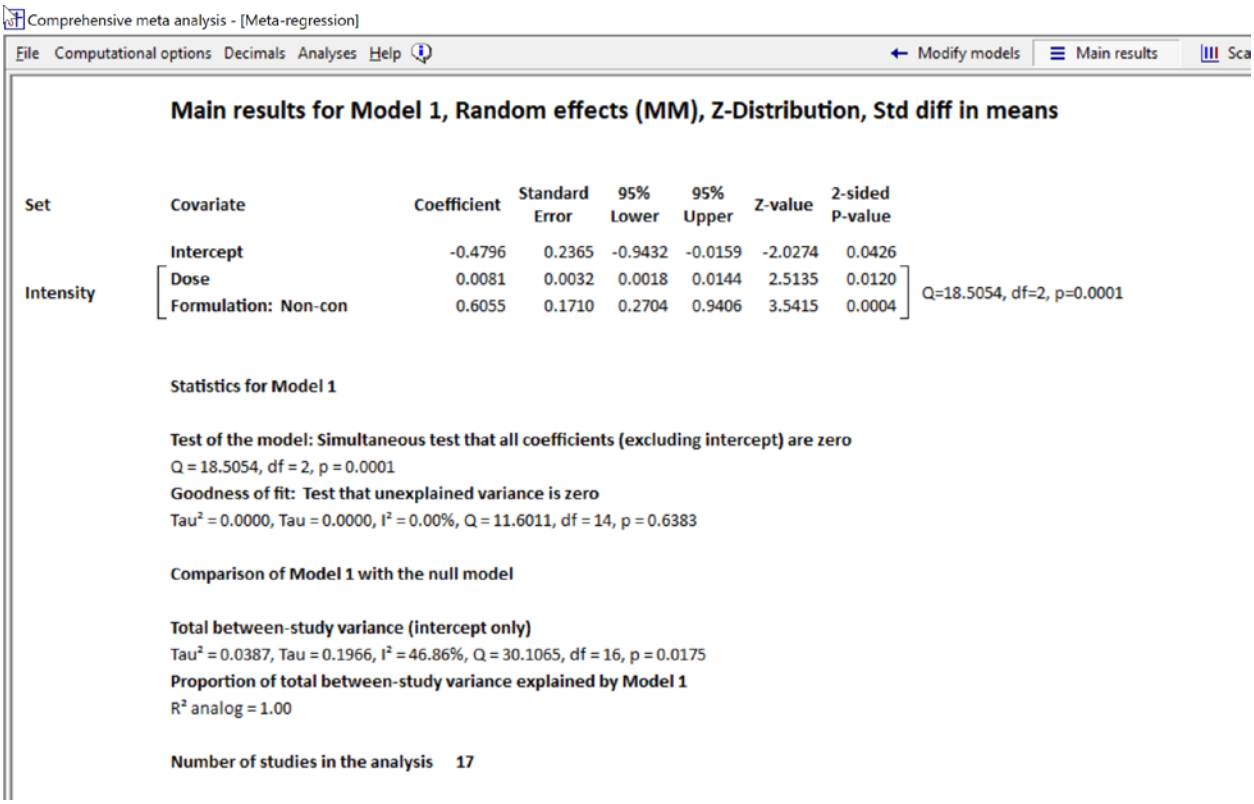
Run the regression

When we control for Formulation, a one-unit increase in Dose is associated with an increase of 0.0081 in the effect size. The Z-value is 2.5135 and the p-value is 0.0120.

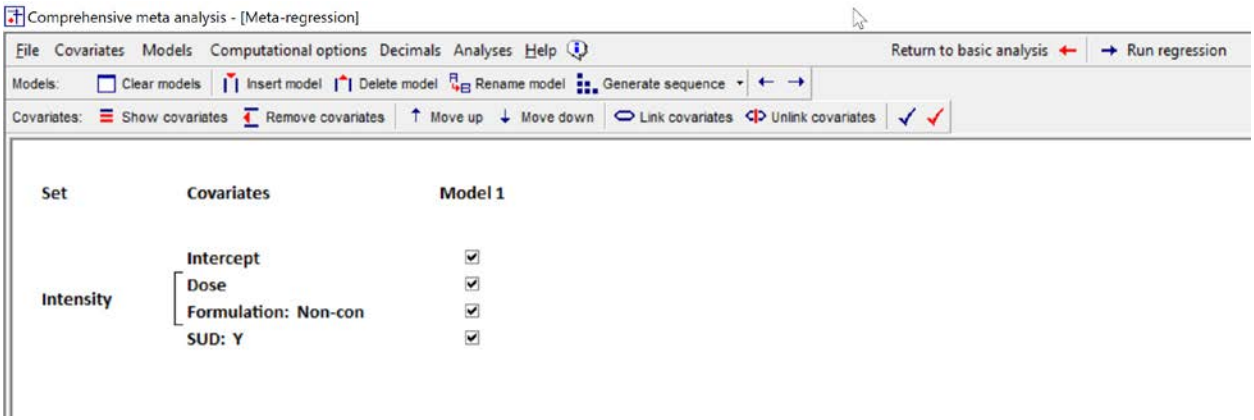
When we control for Dose, a non-continuous formulation is associated with an increase of 0.6055 points in the effect size as compared with a continuous formulation. The Z-value is 3.5415 and the p-value is 0.0004.

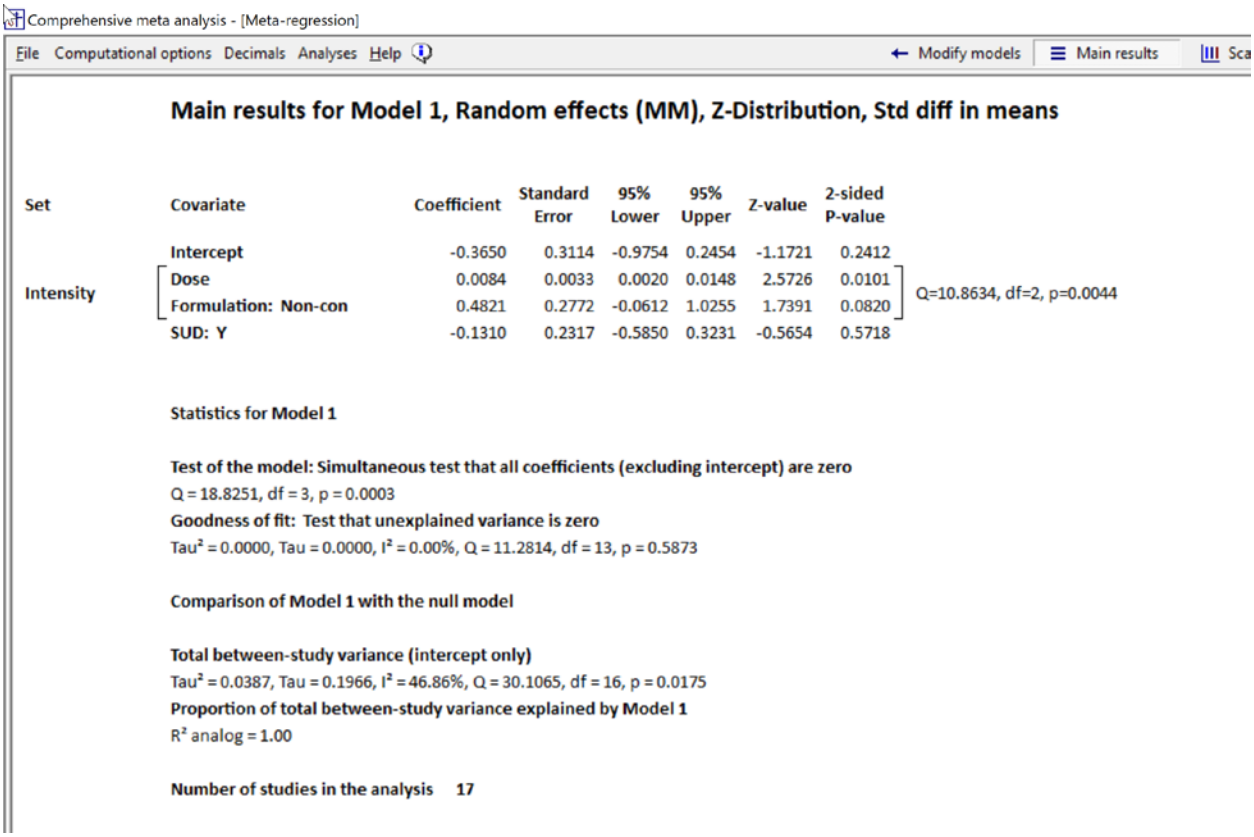
For the two covariates as a set, the Q value is 18.5054 with two degrees of freedom and p=0.0001. We conclude that at least one of the covariates is associated with effect size.

Since the only covariates in the model are Dose and Formulation, we could get the same information from the model. For the model, the Q value is 18.5054 with two degrees of freedom and p=0.0001.



At this point we've established that the intensity of the treatment (as defined above) is related to the effect size. However, we might be concerned that this is due to a confound with SUD. For example, it's possible that studies which used a more intense treatment (higher dose / non-continuous formulation) were more likely to exclude SUD patients. And, it might be the absence of these patients, rather than the intensity of the intervention, that was responsible for the higher effect size. To assess this we can add SUD to the prediction equation





With SUD in the equation, the Q-value for the set is 10.8634 with two degrees of freedom and p=0.0044. So, the relationship between Intensity and effect size cannot be explained (entirely) as a confound with SUD. Put another way, even for studies with the same population (SUD or non-SUD), there will be a relationship between Intensity and effect size.

Note, however, that the unique impact of each covariate is less clear. The relationship between Dose and effect size remains essentially unchanged and statistically significant when we add SUD. By contrast, the relationship between Formulation and effect size is somewhat weaker, and no longer statistically significant.

1. For categorical
2. For others

The data includes a variable called *Dose-C* and another called *Dose-C*^{2}. The former represents the linear relationship between *Dose* and effect size, while the latter represents the curvilinear relationship between *Dose* and effect size. These variables have been centered about the mean as explained in chapter _____.

When we run the analysis we might want to know (a) if there is linear relationship, and (b) if there is a curvilinear aspect to the relationship. We might also want to know (c) if the linear and curvilinear aspects together are related to effect size. To address this third question, we need to create a set.

How to create a Set

Use Dose and Duration

In Figure 88

- Move Dose-C into the model
- Move Dose-C^2 into the model
- Ensure that Dose-C and Dose-C^2 are sequential in the list. To change the sequence, click on one or more covariates and then click [Move Up] or [Move Down] on the toolbar. [A]

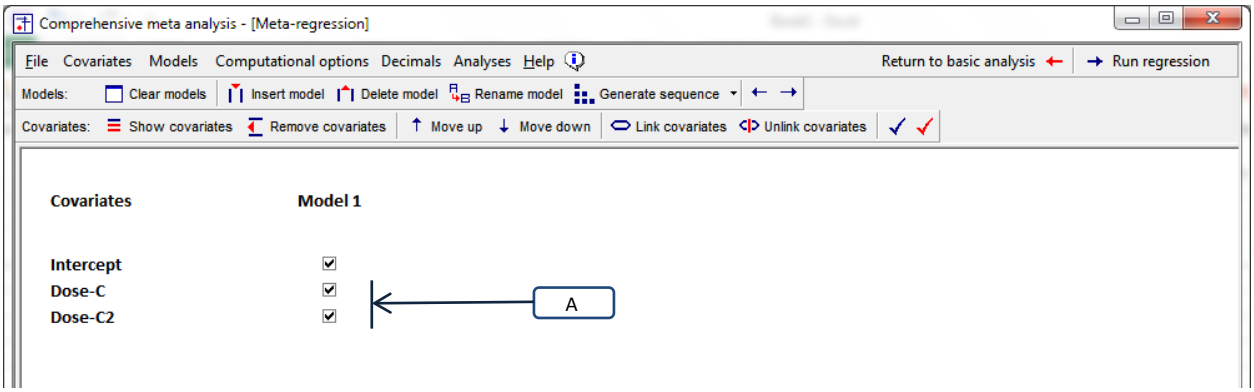


Figure 88 | Setup | Defining a set of covariates

In Figure 89

- Highlight *Dose-C* and *Dose-C*^{2} by pressing [SHIFT] and clicking on these covariate names [B]
- Click [Link Covariates] [C]
- Enter the name Dose Set and click [Ok] [D]
- The program will put brackets around these covariates and identify them as a set [E]

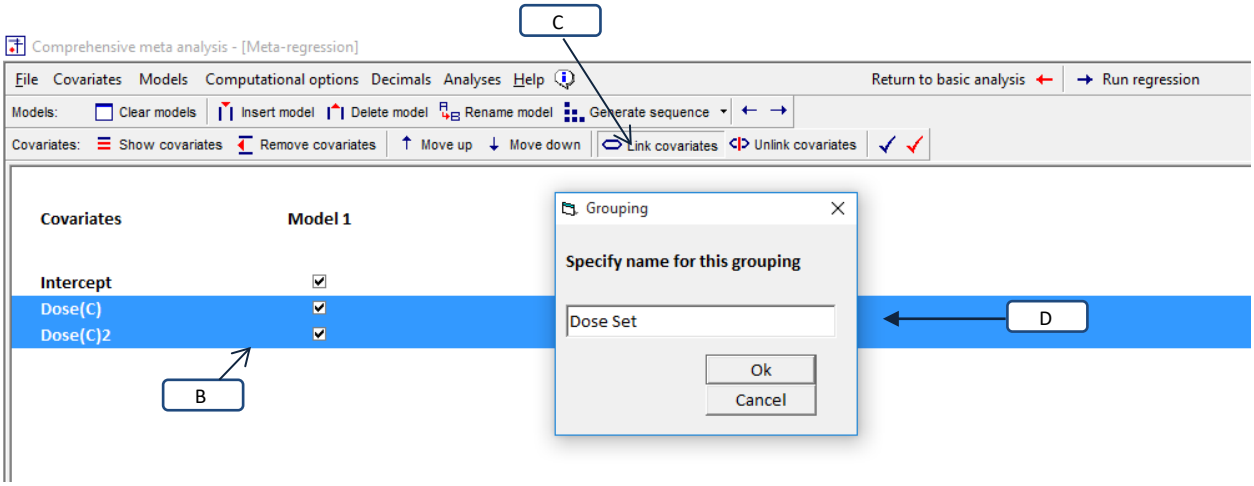


Figure 89 | Setup | Naming a set of covariates

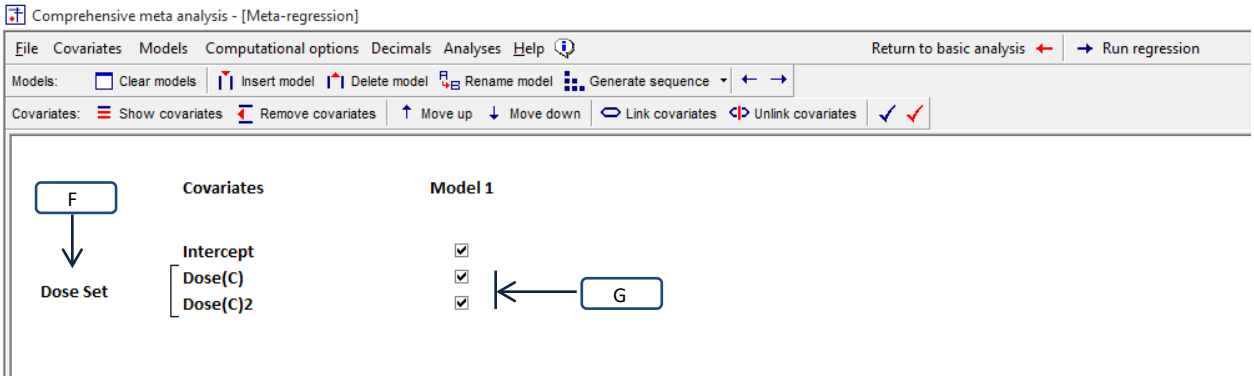


Figure 90 | Setup | Naming a set of covariates

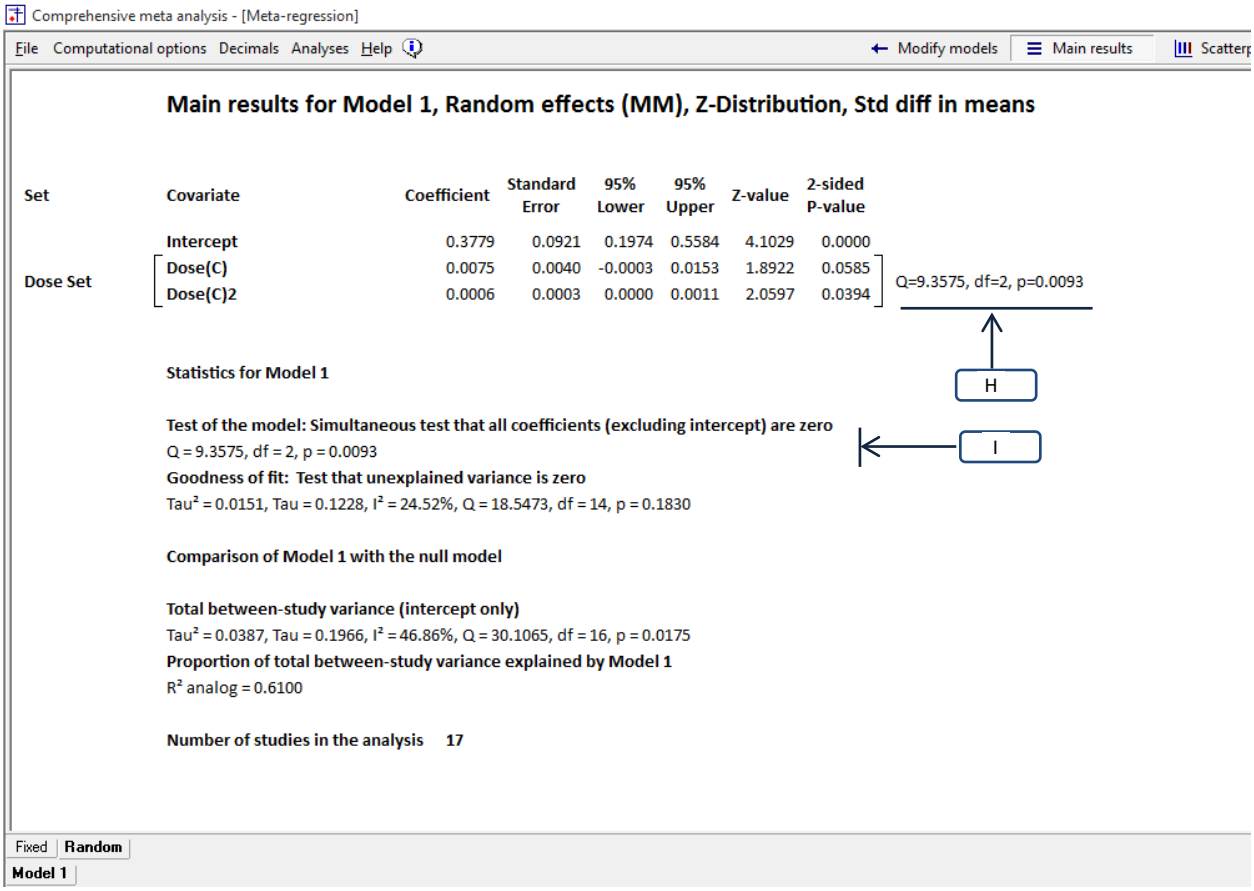


Figure 91 | Regression | Main results | Working with a set of covariates

The results are displayed in Figure 91.

Is Dose-C related to effect size when Dose-C^2 is held constant? (Is there a linear relationship?) The coefficient for Dose-C is 0.0075, with a Z-value of 1.8922 and a p-value of 0.0585

Is Dose-C2 related to effect size when Dose-C is held constant? (Is there a curvilinear linear relationship?) The coefficient for Dose-C2 is 0.0006, with a Z-value of 2.0597 and a p-value of 0.0394.

Are Dose-C and Dose-C^2 (as a set) related to effect size? (Are the linear and curvilinear aspects together able to predict effect size?) This is addressed by a test of the set [H]. The Q-value is 9.3575 with df=2 and p=0.0093

In this example Dose-C and Dose-C^2 are the only covariates in the analysis. Therefore, the test of the set [H] is identical to the test of the model [I]. Therefore, we could have simply used the model and did not need to create a set.

However, consider what happens if we include a third covariate to the model as shown in Figure 92 and Figure 93.

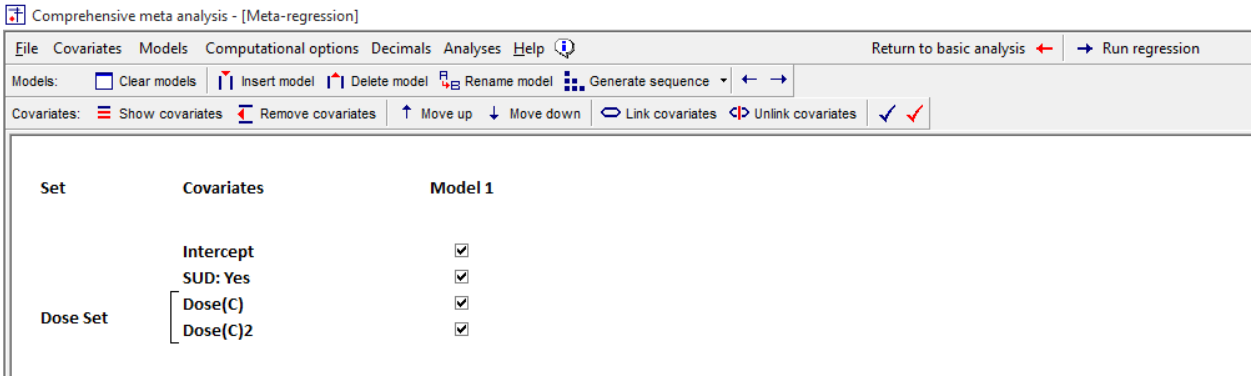


Figure 92

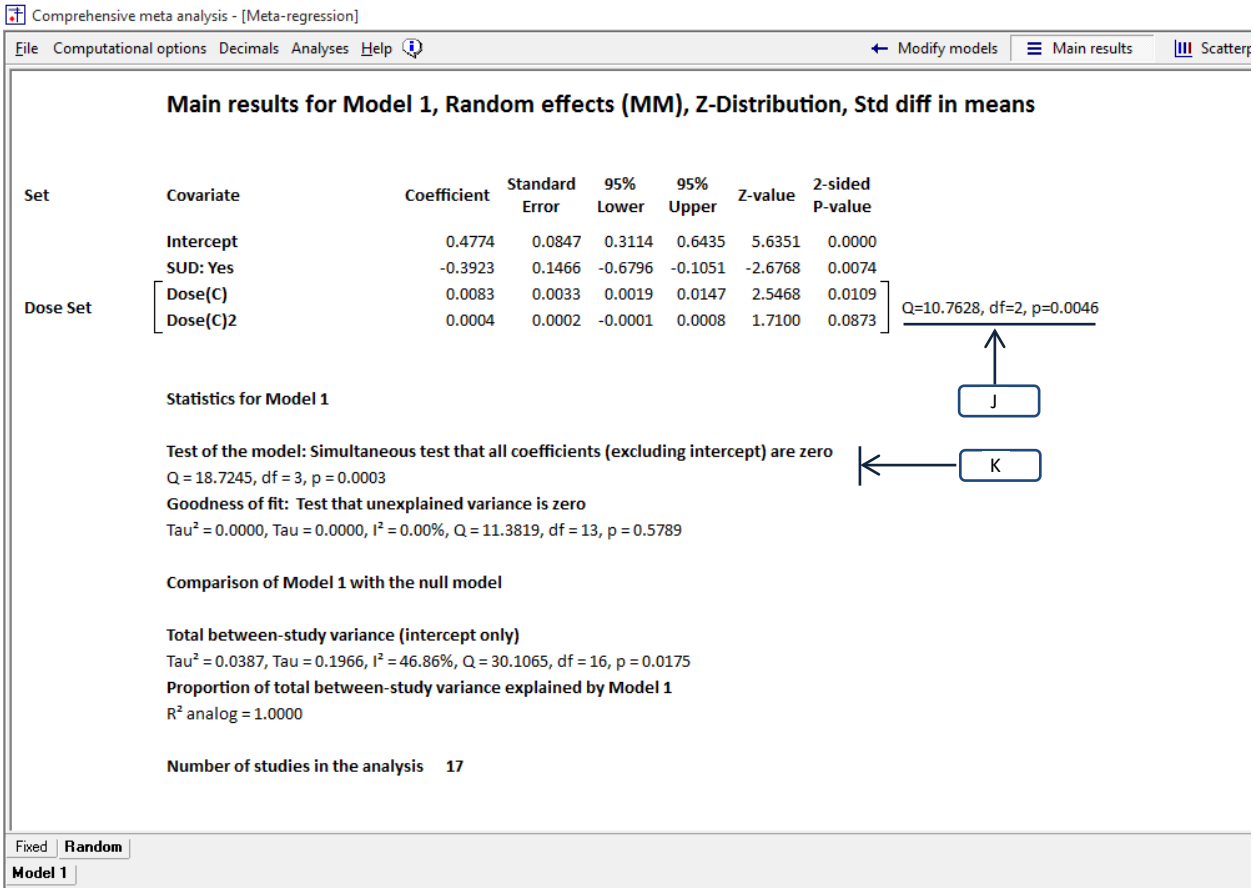


Figure 93

Now, the model includes not only *Dose-C* and *Dose-C2*, but also *SUD*.

The model addresses the impact of *Dose-C*, *Dose-C2*, and *SUD*. The set addresses the impact of *Dose-C* and *Dose-C2* with *SUD* held constant.

- For the model [K] *Q* is 18.7245, df =3, and p=0.0003.
- For the set [J] *Q* is 10.7628, df=2, p=0.0046.

HOW TO REMOVE A SET

To remove a set

- Highlight the set's name [L]
- Click Unlink Covariates [M]

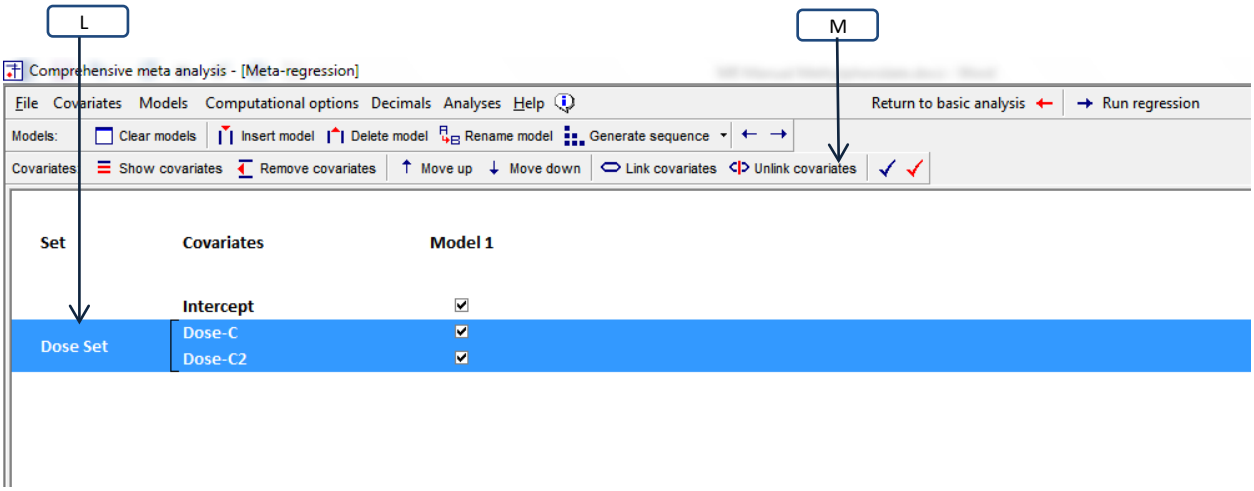


Figure 94 | Main results | Removing a set of covariates

- If you remove the set, the variables will remain in the model, but will no longer be linked.
- You can define more than one set. However, a covariate can only belong to one set at a time.
- To add additional covariates to a set, remove the set, then define a new set

INTERACTIONS

The definition of an interaction is the same in a meta-analysis as it is in a primary study. To wit, suppose we run a regression with two covariates, X_1 and X_2.

- When there is *no* interaction between the covariates, the impact of X_1 on the effect size is constant for all values of X_2 (and vice versa).
- When there *is* an interaction between the covariates, the impact of X_1 on the effect size depends in the magnitude of X_2 (and vice versa).

We'll use an example where X_1 is *Dose*, X_2 is *Duration*, and X_3 is the interaction of X_1 and X_2.

When we include X_1 and X_2 in the prediction model but don't include X_3, we are assuming that the impact of Dose is constant for all values of Duration (and vice-versa), and are forcing the prediction lines to be parallel as in Figure 95. In this case

- We ask "What's the relationship between Dose and effect size" without specifying a time frame, because we assume that the relationship between Dose and effect size (the slope of the regression line) is the same for all Durations. In Figure 95, the difference in effect size for a dose of 81 vs. a dose of 31 is ____ points regardless of Duration. If the p-value for Dose is statistically significant, we would conclude that Dose is related to effect size for any Duration.

- We ask "What's the relationship between Duration and effect size" without specifying a Dose, because we assume that the relationship between Duration and effect size (the distance between the regression lines) is the same at all Doses. In Figure 95, the difference in effect size for a Duration of 10 days vs. 90 days is ____ points regardless of dose. If the p-value for Duration is statistically significant, we would conclude that Duration is related to effect size for any Dose.

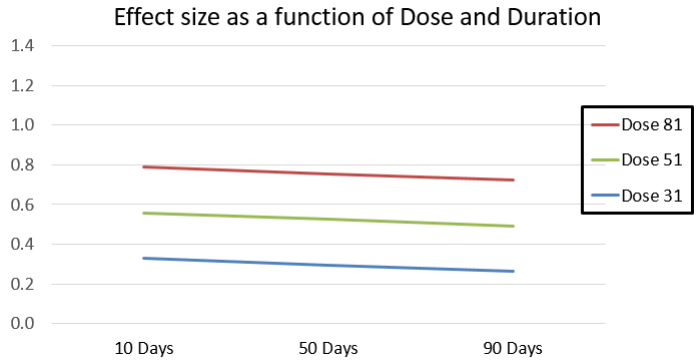


Figure 95

To be clear, the fact that the lines are parallel to each other reflects *our decision* to omit the interaction from the regression term. Without the interaction term in the equation, the lines *must* be parallel.

When we do include the interaction term we allow that the impact of *Dose* may depend on the study duration (and vice-versa), and we allow the lines to converge, as in Figure 96. In this example,

Rather than ask "What's the relationship between Dose and effect size", we would ask "What's the relationship between Dose and effect size for studies with a Duration of *X* days". Since the relationship between Dose and effect size varies by study duration, it follows that the *p*-value for Dose will depend on the study duration. When we report a *p*-value, this *p*-value is only valid for studies of a specific duration. The way regression works, the relationship between Dose and effect size will be evaluated for the case where the study Duration is zero.

The same idea applies also to the other covariate. Rather than ask "What's the relationship between Duration and effect size", we would ask "What's the relationship between Duration and effect size for studies with a Dose of *X* units". Since the relationship between Duration and effect size varies by Dose, it follows that the *p*-value for Duration will depend on the Dose. When we report a *p*-value, this *p*-value is only valid for studies that employed a specific dose. The way regression works, the relationship between Duration and effect size will be evaluated for the case where the study Dose is zero.

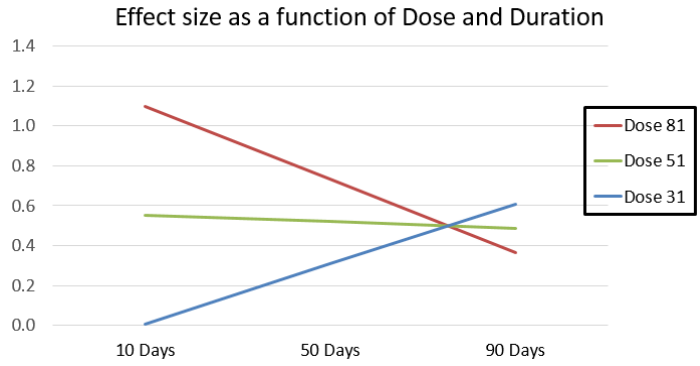


Figure 96

So, we would assess the impact of Dose for the case where Duration is zero, but this is not terribly informative since a study of zero Duration is of no interest. It would make more sense to assess the impact of Dose for a study where the Duration is at some meaningful value – for example, the mean, or some other Duration of interest.

Similarly, we would assess the impact of Duration for the case where Dose is zero, but this is not terribly informative since a study with a Dose of zero is of no interest. It would make more sense to assess the impact of Duration for a study where the Dose is at some meaningful value – for example, the mean, or some other Dose of interest.

The way that we accomplish this goal is by a procedure called centering. This simply means that we shift the scale of a covariate, so that some value is assigned a score of zero, and then all other values are centered about that score.

For example, suppose that we had a total of five studies, with dosages of 30, 40, 50, 60, 70. To center about the Dose of 50 we would subtract 50 from each, to yield values of -20, -10, 0, +10, and +20. We might call the centered variable Dose-C.

Consider what happens if we run the regression using Dose-C, Duration, and the interaction of Dose-C with Dose. The impact of Duration would be evaluated for the case where Dose-C is 0. In fact, this is the case where Dose is 50. So, we would be assessing the impact of Duration for a case that we care about.

It's common to center a variable about its mean (as in this example) but we could also center it about some other value. For example, if we chose to center Dose about a value of 30, we would assess the impact of Duration for a Dose of 30.

The same idea applies to Duration. We could center Duration about the mean duration, to create a variable called Duration-C. Then, the impact of Dose would be evaluated for the case where Duration-C is zero and (it follows) the Duration is the mean Duration.

Typically, we would center both Duration and Dose. Then, we would evaluate the impact of each covariate for a meaningful value of the other.

The same idea holds true for categorical variables.

Suppose we have a variable called SUD which is coded 1 for studies that include SUD patients, and 0 for studies that exclude SUD patients. Suppose further that we run a regression using Dose, SUD, and the interaction between them. In this case we would evaluate the impact of Dose for studies that are coded 0 on SUD, which is studies that exclude these patients. If that's our intent, then this will work well. However, there are other options, including the following.

We could create another variable called Clean, coded 1 for studies that exclude SUD patients, and 0 for studies that include them. In this case we would evaluate the impact of Dose for studies that include SUD patients.

In this sample of 17 studies, 4 include SUD (coded 1) and 13 exclude SUD (coded 0) so the mean on this variable is 4/17 or 0.24. We could create a variable called SUD-C which is centered on 0.24. This would assess the impact of Dose for a "typical" study.

Along these lines, we could assign a code of 1 to studies that include SUD patients and a code of -1 to studies that exclude SUD patients. Then we would evaluate the impact of Dose for studies with a code of 0 on SUD. This would be a hypothetical study that was neutral on SUD.

The decision to adopt any given approach will depend on the questions that we want to address. The key point is that

Decisions about centering X1 will affect the results for X2.
Decisions about centering X2 will affect the results for X1.

Decisions about centering either covariate will not affect results for the interaction.

The problem is that these comparisons would not be terribly informative

That is, we don't really want to know

MAKE THIS A SECTION, AND DO FOR (1) CONTINUOUS (2) CATEGORICAL (3) CURVILINEAR

Centering

To center a variable means to re-scale the variable about some value. For example, if we had five studies with dosages of 30, 40, 50, 60, 70, we could subtract 50 from each, to yield values of -20, -10, 0, +10, and +20. These values would be centered about 50.

This is useful when we're working with interactions for the following reason. When we include X_1, X_2, and X_3 (the interaction) in the prediction model, then the impact of X_1 is tested for the case where X_2 is zero, and the impact of X_2 is tested for the case where X_1 is zero.

With this in mind, consider Figure 96.

- If we enter *Dose* into the equation in the original metric we would be testing the impact of *Duration* for the (meaningless) case where *Dose* is zero. Instead, we create a variable called *Dose-C* (centered) where a Dose of 56 is entered as zero, and all other doses are assigned scores relative to this one. Now, the impact of *Duration* will be tested for the case where *Dose-C* is zero and (it follows) *Dose* is 56.

- If we enter *Duration* into the equation in the original metric we would be testing the impact of *Dose* for the (impossible) case where *Duration* is zero. Instead, we create a variable called *Duration-C* (centered) where a *Duration* of 50 is entered as zero, and all other durations are assigned scores relative to this one. Now, the impact of *Dose* will be tested for the case where *Duration-C* is zero and (it follows) *Duration* is 50.

While we often center on the mean, we can actually center on any value. The key point is that we want the value of zero to represent a meaningful level of the variable. IN THIS EXAMPLE

When X1 is categorical, and dummy coded 0 or 1, we could center this variable on the mean, in which case X2 would be evaluated for the "typical" value of X1. Alternatively, we could choose not to center X1. In this case X2 would be evaluated for studies where X1 was coded 0.

Note that the use of centering only matters if we include the interaction. If we include only main effects in the regression, then the *p*-value for *Duration* is the same whether or not we center *Dose*. And the *p*-value for *Dose* is the same whether or not we center *Duration*.

The same rules apply for interactions involving categorical variables, continuous variables, or combinations of the two types. For clarity, we present an example for each of three cases. These are

- The interaction of two categorical covariates
- The interaction of a categorical covariate with a continuous covariate
- The interaction of two continuous covariates

NEW SECTION

Centering is also important if we want to assess the impact of curvilinear relationships. For example, suppose that we want to see if the relationship between Dose and effect size has a curvilinear component. For this purpose we need to enter both Dose and Dose ^{2} as covariates.

If we don't center Dose, then these two covariates will be highly correlated with each other, and it will be difficult to disentangle the linear from the curvilinear components. By contrast, if we center Dose and square the centered value, the correlation between the two will be low, and we will be able to identify the unique impact of each.

In fact, the case of a curvilinear relationship can be seen as an example of the interaction between two continuous variable. We said that an interaction exists when the impact of X1 depends on the value of X2. Here, the impact of X1 depends on the value of X1. If we ask "What happens if we increase Dose by 10 units" the answer is "It depends". For some values of Dose, a 10-unit increase will have one effect. For other values of Dose, a 10-unit increase will have another effect.

If the relationship between Dose and effect size is curvilinear, then the impact of X1 (Dose) depends on the level of X2 (Dose

The program will not automatically create centered variables, nor variables for the interaction. While it is possible to enter these variables manually, it's usually easier to copy the original variables to Excel™, create the new variables, and then copy these back into CMA. (See appendix)

INTERACTION OF TWO CATEGORICAL COVARIATES

The original data set includes the covariates Dose and Duration, both of which are continuous. For purposes of this discussion we need two categorical covariates, and we create them by dichotomizing the variables to create Dose (I) and Duration (I) where the (I) stands for "Integer". (see Appendix 5: Creating variables for interactions).

- Long (I) is coded 0 for short Duration (<40), and 1 for long Duration
- High (I) is coded 0 for low dose (< 52), and 1 for high dose
- High(I) x Long(I) is the interaction

Figure 97 shows the model, Figure 98 shows the main results, and Figure 99 shows the corresponding plot. While the program can compute the statistics for interactions it cannot *plot* these interactions, and therefore these plots were created in Excel™ (see appendix).

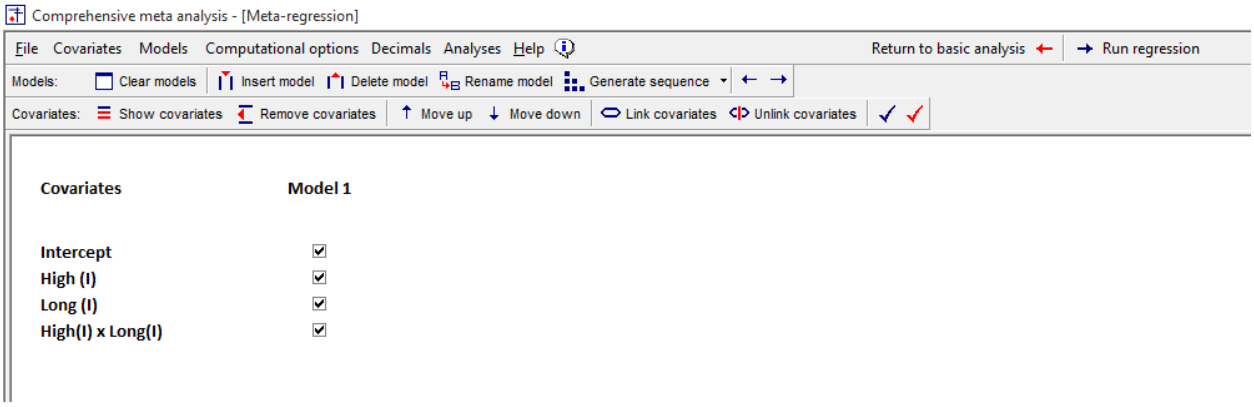


Figure 97 | Setup | Interaction of two categorical covariates

Comprehensive meta analysis - [Meta-regression]

File Computational options Decimals Analyses Help Modify models Main results Scatterp

Main results for Model 1, Random effects (MM), Z-Distribution, Std diff in means

Covariate	Coefficient	Standard Error	95% Lower	95% Upper	Z-value	2-sided P-value	
Intercept	0.4288	0.1155	0.2024	0.6552	3.7122	0.0002	A
High (I)	0.4801	0.2655	-0.0402	1.0005	1.8084	0.0705	B
Long (I)	-0.1379	0.2088	-0.5471	0.2714	-0.6602	0.5091	C
High(I) x Long(I)	-0.1835	0.3393	-0.8485	0.4815	-0.5408	0.5886	D

Statistics for Model 1

Test of the model: Simultaneous test that all coefficients (excluding intercept) are zero E

Q = 5.2787, df = 3, p = 0.1525

Goodness of fit: Test that unexplained variance is zero

Tau^2 = 0.0342, Tau = 0.1849, I^2 = 41.26%, Q = 22.1302, df = 13, p = 0.0534

Comparison of Model 1 with the null model

Total between-study variance (intercept only) F

Tau^2 = 0.0387, Tau = 0.1966, I^2 = 46.86%, Q = 30.1065, df = 16, p = 0.0175

Proportion of total between-study variance explained by Model 1

R^2 analog = 0.1157

Number of studies in the analysis 17

Figure 98 | Main results | Interaction of two categorical covariates

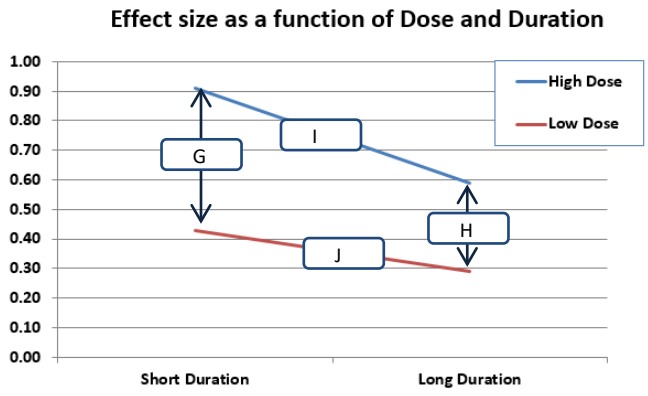


Figure 99

Does the drug's impact vary as a function of *High-Dose vs. Low-Dose?*

When we report the impact of Dose we do so for the studies that have been coded zero for Duration, which in this case is the short-duration studies. In Figure 98, the coefficient for Dose (I) is 0.4801. This tells us that for Short-Duration studies, the effect size for High-Dose studies is 0.4801 higher than for Low-Dose studies. In Figure 99, this corresponds to arrow [G]. For short-duration studies, the predicted effect size is 0.4288 for Low-Dose and 0.9089 for High-Dose. The difference between these is equal to the coefficient for Dose (I), which is 0.4801 (Figure 98 B), and the *p*-value of 0.0705 corresponds to the effect represented by this arrow.

Does the impact of the drug differ for Short-Duration studies vs. Long-Duration studies?

When we report the impact of Duration we do so for the studies that have been coded zero for Dose (I), which in this case is the Low-Dose studies. In Figure 98, the coefficient for Long (I) is −0.1379. This tells us that for Low-Dose studies, the effect size for Long-Duration studies is 0.1379 lower than for Short-Duration studies.

In Figure 99, this corresponds to line [J]. For Low-Dose studies, the predicted effect size is 0.4288 for Short-Duration and 0.2909 for Long-Duration. The difference between these is equal to the coefficient for Long (I), which is −0.1379 (Figure 98 C), and the *p*-value of 0.5090 corresponds to the effect represented by this difference.

Dose x Duration

Does the relationship between Dose and effect size vary by Duration? In Figure 99 this would mean that arrow [G] is a different height than the arrow [H]. Arrow [G] shows that the impact of Dose for Short-Duration studies is 0.4801, whereas arrow [H] shows that the impact of Dose for Long-Duration studies is 0.2966. The difference between these two (0.1835) is the interaction of Dose by Duration. This is also the coefficient for the interaction in Figure 98.

Does the relationship between Duration and effect size vary by Dose? In Figure 99 this would mean that the slope for Duration in High-Dose studies [I] is not the same as the slope for Duration in Low-Dose studies [J]. Slope [I] shows that the impact of Duration for High-Dose studies is 0.3214 whereas slope [B] shows that the impact of Duration for Low-Dose studies is 0.1379. The difference between these two (0.1835) is the interaction of Dose by Duration. This is also the coefficient for the interaction in Figure 98.

These two questions are functionally identical. Asking if slopes [I] and [J] are identical is functionally the same as asking if arrows [G] and [H] are the same height. It follows that the same *p*-value applies to both questions. The *p*-value is the p-value for the interaction, which is reported as 0.5886 in Figure 98 [D].

The full set

Is there a relationship between Duration (Short vs. Long), Dose (Low vs. High), and the interaction (as a set) and the effect size? Since these three are the only covariates in the model, this is addressed by a test of the full model, (Figure 98 E).

- The *p*-value of 0.1525 does not allow us to reject the null hypothesis. We cannot conclude that some of the covariates are related to effect size.

- The R^2 analog [F] is 0.1157, which tells us that around 12% of the initial between-study variance in effect sizes can be explained by this combination of covariates. However, given the non-significant p-value (0.1525) the true value of R^2 could be 0.00.

- Thus, the analysis does not provide evidence that the full model is able to explain at any of the variance in effect size.

Alternatively, could have created a set

Notes

If we wanted to test for H, not G –

If I, not T

INTERACTION OF A CATEGORICAL COVARIATE WITH A CONTINUOUS COVARIATE

The original data set includes the covariates Dose and Duration, both of which are continuous. For purposes of this discussion we need a categorical covariate, and we create it by dichotomizing Dose (see Appendix 5: Creating variables for interactions).

In this analyses we will be using the following covariates

- Duration-C is the Duration, centered on the mean duration (49.96)
- High(I) is coded 0 for low dose (< 52), and 1 for high dose
- Duration-C x High(I) is the interaction

Figure 100 shows the model, Figure 101 shows the main results, and Figure 102 shows the corresponding plot. While the program can compute the statistics for interactions it cannot *plot* these interactions, and therefore the plot (Figure 102) was created in Excel™ (see appendix).

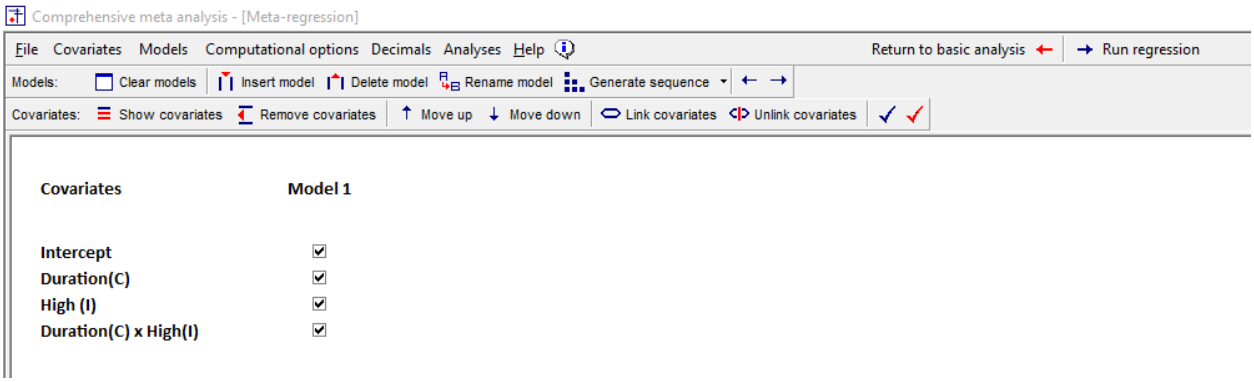


Figure 100 | Setup | Interaction of categorical and continuous covariates

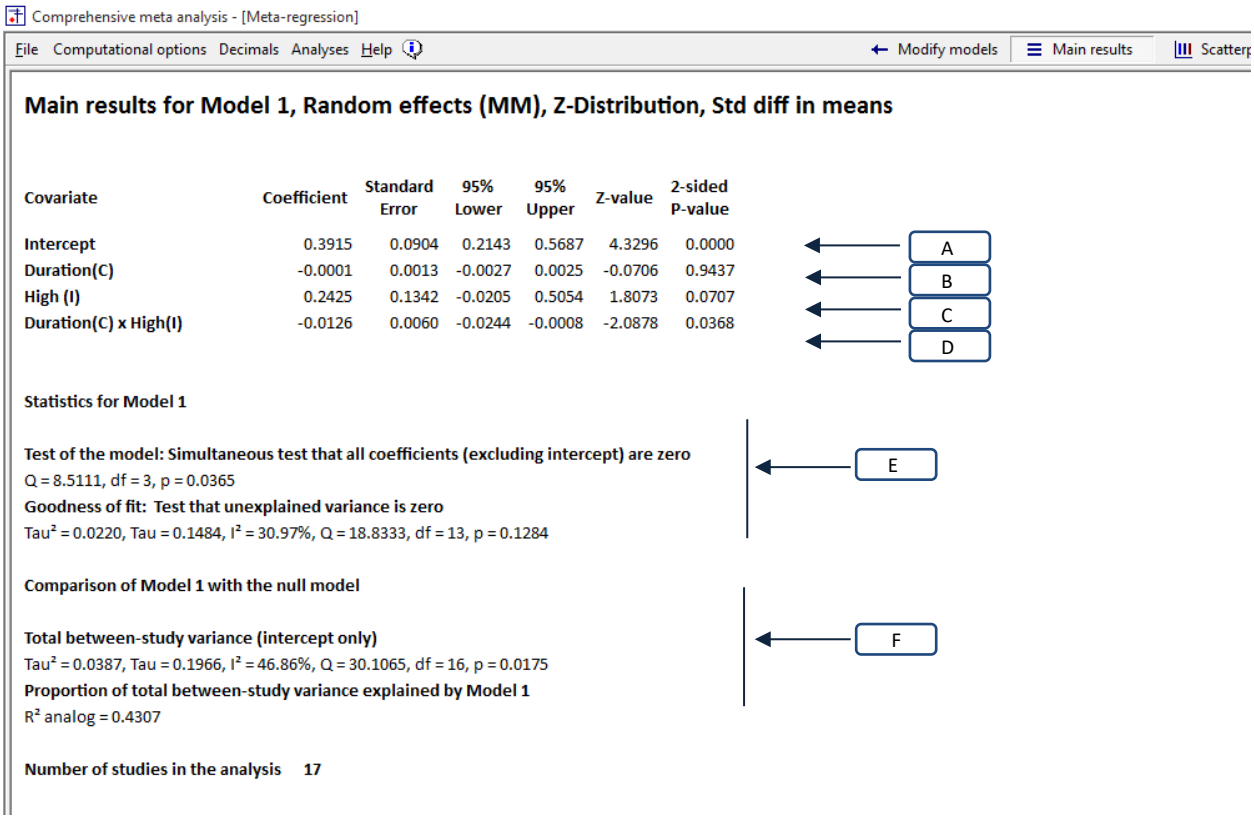


Figure 101 | Main results | Interaction of categorical and continuous covariates

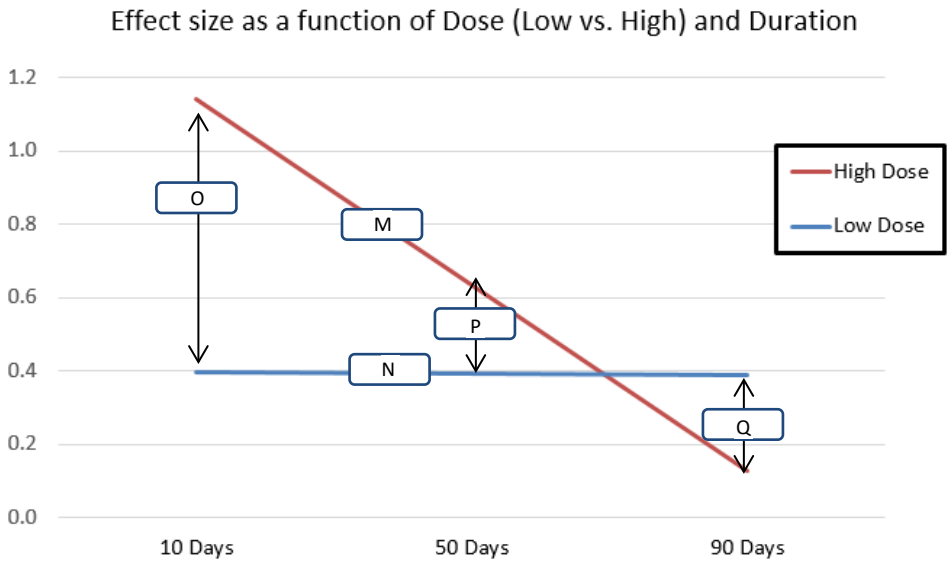


Figure 102 | Plot | Interaction of two continuous covariates

Does the drug's impact differ as a function of *High-Dose vs. Low-Dose?*

The regression reports the impact of dose for the case where Days-C is zero. Since we've centered at the mean duration of 50 days, this corresponds to a study of 50 days. In Figure 101 [C], the coefficient for High (I) is 0.2425. This tells us that for studies with a duration of 50 days, the effect size for High-Dose studies is 0.2425 higher than for Low-Dose studies.

In Figure 102, this corresponds to arrow [P]. For 50-day studies, the predicted effect size is 0.63 for High-Dose and 0.39 for Low-Dose. The difference between these is equal to the coefficient for High (I), which is 0.24 (Figure 98 B), and the *p*-value of 0.0707 corresponds to the effect represented by this arrow.

Duration

Is the drug's impact related to study duration?

When we report the impact of Duration we do so for the studies that have been coded zero for High (I), which in this case is the Low-Dose studies. In Figure 101 [B], the coefficient for *Duration (C)* is −0.0001. This tells us that for Low-Dose studies, the effect size drops by .0001 for every day of study duration.

In Figure 102, this corresponds to line [N]. For Low-Dose studies, the predicted effect size is 0.40 for a 10-day study vs. .39 for a 90-day study. This difference (0.01) is equal to the coefficient for Duration (0.0001) times the number of days between the plotted points (80). The coefficient for duration is reported as .0001 in _____. The corresponding p-value of 0.9437 is not affected by the time-points in the plot.

Interaction of Duration x Dose (Low vs. High)

Does the relationship between Dose and effect size vary by Duration (High vs. Low)? In Figure 102 this would mean that arrow [O] is a different height than the arrow [Q]. Arrow [O] shows that the impact of Dose for 10-day studies is 0.7450 in favor of high-dose, whereas arrow [P] shows that the impact of Dose for 90-day studies is 0.2600 in favor of low-dose. The difference between these two (1.0050) is the interaction of Dose by Duration. This is also the coefficient for the interaction in Figure 101. TIMES 50?

Does the relationship between Duration and effect size vary by Dose? In Figure 102 this would mean that the slope for Duration in High-Dose studies [M] is not the same as the slope for Duration in Low-Dose studies [N]. Slope [M] shows that the impact of Duration for High-Dose studies is a decrease of 1.0124 for an 80-day increase in duration, whereas slope [N] shows that the impact of Duration for Low-Dose studies is a decrease of 0.0074 over the same change in duration. The difference between these two (1.00496) is the interaction of Dose by Duration. This is also the coefficient for the interaction in Figure 101. TIMES 50?

These two questions are functionally identical. Asking if slopes [M] and [N] are identical is functionally the same as asking if arrows [O] and [Q] are the same height. It follows that the same *p*-value applies to both questions. The *p*-value for this interaction, is reported as 0.0368 in Figure 98 [D].

The full set

Is there a relationship between Duration, Dose (Low vs. High), and the interaction (as a set) and the effect size? Since these are the only covariates in the model, this is addressed by a test of the full model [Figure 101 E].

- The *p*-value of 0.0365 allows us to reject the null hypothesis that none of the covariates is related to effect size.

- The R^2 analog [Figure 101 F] is 0.4307, which tells us that 43% of the initial between-study variance in effect sizes can be explained by this combination of covariates.

- Thus, we conclude that the full model (Duration, Dose (Low vs. High), and the interaction) is able to explain at least some of the variance in effect size.

IF WE WANTED TO TEST FOR Q

IF WE WANTED TO TEST FOR DURATION =

INTERACTION OF TWO CONTINUOUS COVARIATES

In this example we assess the impact of Dose, Duration, and the interaction between them. It's generally a good idea to center variables when including the interaction in the prediction equation, and that is the practice we follow here.

For instructions on creating the data set used here, see Appendix 5: Creating variables for interactions

- Dose-C is the Dose, centered on the mean dose (55.68)
- Days-C is the Duration, centered on the mean duration (49.96)
- Dose-C x Duration-C is the interaction

Figure 103 shows the model, Figure 104 shows the main results, and Figure 105 shows the corresponding plot. While the program can compute the statistics for interactions it cannot *plot* these interactions, and therefore Figure 105 was created in Excel™ (see Plotting the interaction of two continuous covariates).

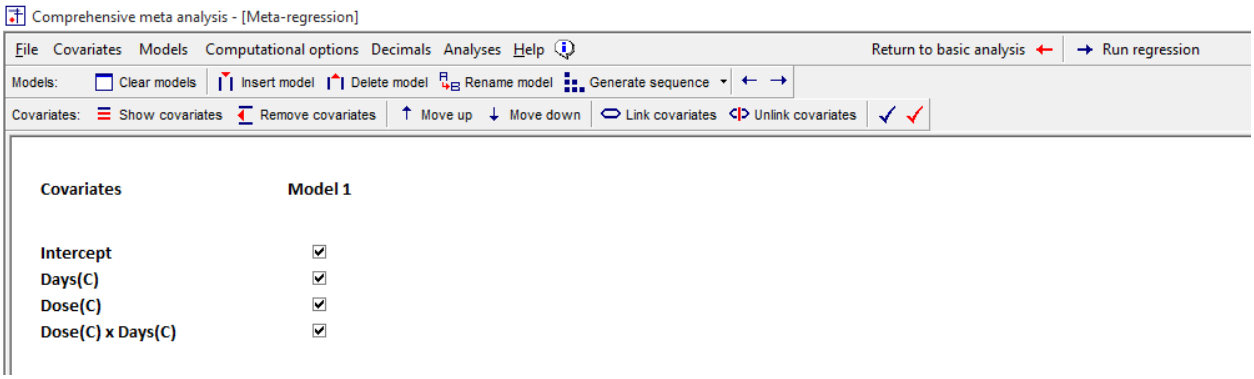


Figure 103 | Setup | Interaction of two continuous covariates

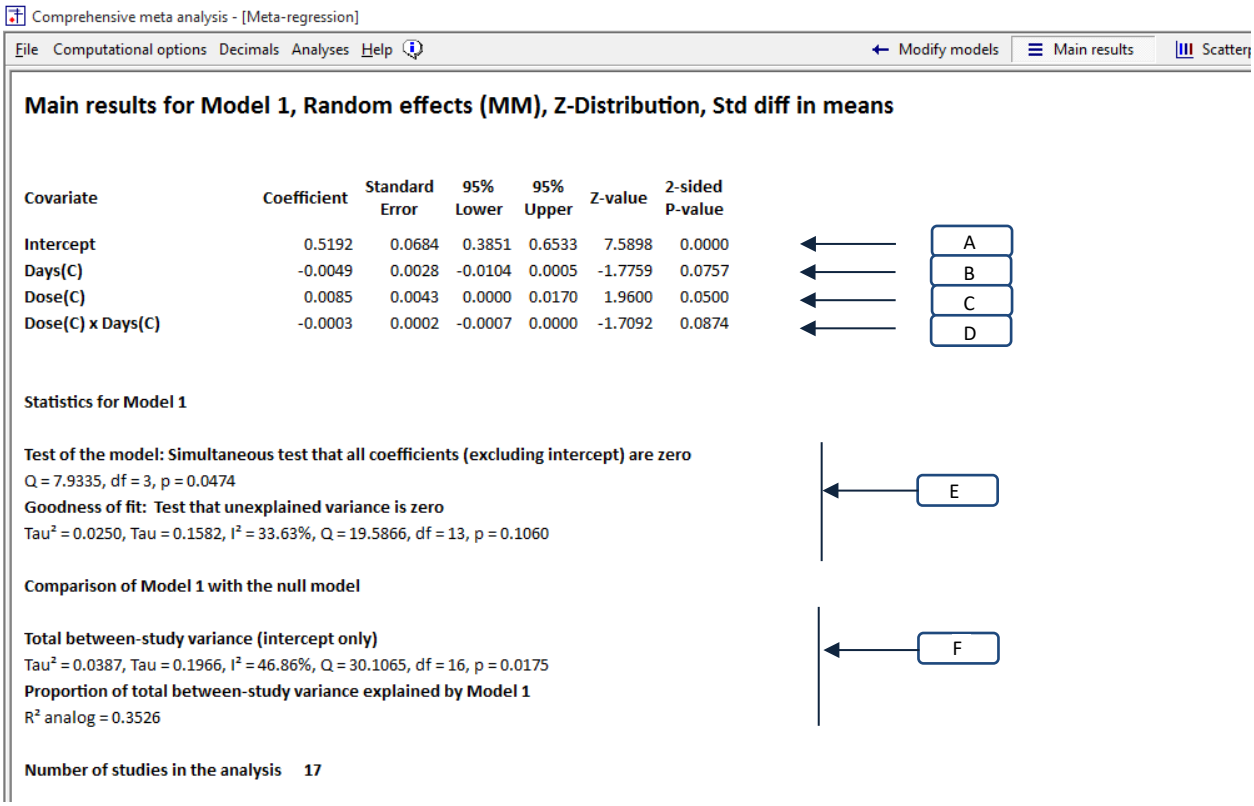


Figure 104 | Main results | Interaction of two continuous covariates

Dose

Does the drug's impact differ as a function of *Dose*?

MOVE UP

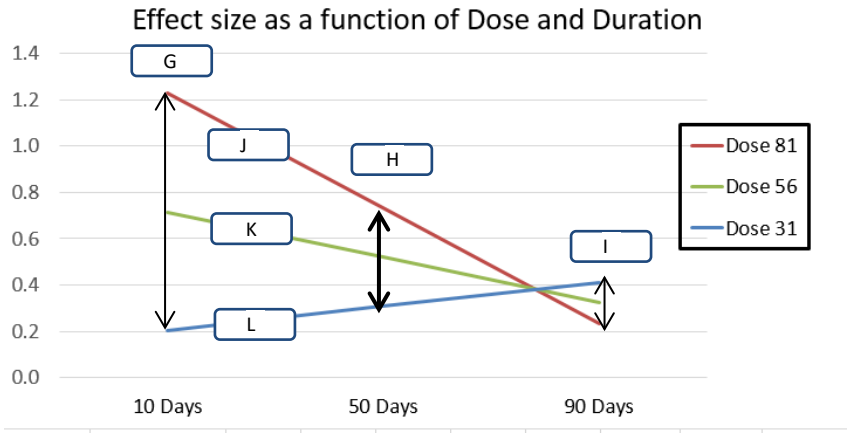


Figure 105 | Plot | Interaction of two continuous covariates

The regression reports the impact of dose for the case where Days-C is zero. Since we've centered at the mean duration of 50 days, this corresponds to a study of 50 days. In Figure 104, the coefficient for Dose (C) is 0.0085. This tells us that for studies with a duration of 50 days, as the dosage increases by 1 unit the effect size increases by .0085.

In Figure 105 this corresponds to arrow [H]. For 50-day studies, the predicted effect size is 0.35 for Dose of 81 units vs. 0.70 for a dose of 31 units. The difference between these is equal to the coefficient for Dose, which is 0.0085 (Figure 104 C) times the difference between the two doses (50 units). The *p*-value of 0.0500 corresponds to the effect represented by this arrow.

While the *p*-value for *Dose* only applies to studies that ran for 50 days, the test does not depend on the fact that we chose to plot effects for doses of 31 and 81. The test depends on the change per unit of dose, and not on the specific values plotted.

Duration

Is the drug's impact related to study duration?

The regression reports the impact of Duration for the case where Dose-C is zero. Since we've centered Dose at the mean Dose of 56, this corresponds to a study where the Dose is 56. In Figure 104 [B], the coefficient for *Duration (C)* is −0.0049. This tells us that for studies where the Dose is 56, the effect size drops by .0049 for every day of study duration, and the p-value is 0.0757.

In this corresponds to line [K]. For studies where the Dose is 56, the predicted effect size is 0.70 for a 10-day study vs. .30 for a 90-day study. This difference (0.40) is equal to the coefficient for Duration (C) (−0.0049) times the number of days between the plotted points (80). The coefficient for duration is reported as −0.0049 in Figure 104 and the corresponding p-value is 0.0757.

While the p-value for *Duration (C)* only applies to studies that employed a dose of 56, the test does not depend on the specific values of Duration that we chose to plot (10 to 90). The test depends on the change per day, and not on the specific values plotted.

Interaction of Dose x Duration

Does the relationship between Dose and effect size vary by Duration? In Figure 105, this would mean that arrow [O] is a different height than the arrow [P]. Arrow [O] shows that the impact of Dose (31 units vs. 81 units) for 10-day studies is 1.02 in favor of the higher dose, whereas arrow [P] shows that the impact of Dose for 90-day studies is 0.18 in favor of the lower dose. The difference between these two effects reflects the interaction.

Does the relationship between Duration and effect size vary by Dose? In In Figure 105, this would mean that the slope for Duration in the 81-Dose studies [M] is not the same as the slope for Duration in the 31-Dose studies [N]. Slope [M] shows that for 81-units studies, the change from 10 days to 90 days is associated with a drop of 0.992 in effect size. Slope [N] shows that for 31-unit studies, the change from 10 days to 90 days is associated with an increase of 0.208 in effect size. The difference between these two effects reflects the interaction.

These two questions are functionally identical. Asking if slopes [I] and [J] are identical is functionally the same as asking if arrows [G] and [H] are the same height. It follows that the same coefficient (−0.0003) and *p*-value (0.0368) applies to both in Figure 104 [D].

As noted, the coefficient for the interaction is −0.0003. If we multiply −0.0003 by 50 (the difference between 31 units and 81 units) and then by 80 (the difference between 10 days and 90 days) we get −1.2000. This is difference between the impact of 50 dose units at 10 days (O) vs the impact of 50 dose units at 90 days (P) is −1.2000. It is also the difference between the impact of 80 days for the high-dose studies [M] vs the impact of 80 days for the low-dose studies [N].

IF WE WANTED TO TEST FOR DOSE = OR FOR DURATION =
CENTERING AFFECTS FIRST ORDER EFFECTS, NOT INTERACTION

The full set

Is there a relationship between Duration-C, Dose-C, and the interaction (as a set) and the effect size? Since these are the only covariates in the model, this is addressed by a test of the full model, as shown in Figure 104 [E].

- The *p*-value of 0.0474 allows us to reject the null hypothesis that none of the covariates is related to effect size (and conclude that at least one of them probably is related to effect size).

- The R^2 analog [F] is 0.3526, which tells us that 35.26% of the initial between-study variance in effect sizes can be explained by this combination of covariates.

- Thus, we conclude that the full model (Duration, Dose, and the interaction between them) is able to explain at least some of the variance in effect size.

CURVILINEAR RELATIONSHIPS

Earlier, we established that there is a linear relationship between Dose and effect size. Suppose we want to test the hypothesis that the relationship between Dose and effect size is actually curvilinear – for example, that the drug's impact is relatively constant as we move from a dose of 30 to 60, but then increases as we move from 60 to 80. Or, that the drug's impact increases sharply as we move from a Dose of 30 to 60, but is relatively unchanged beyond that point. Or, that drug is most effective (or least effective) at moderate values as compared with low or high values.

A curvilinear relationship can be seen as a kind of interaction, and that is the approach we take here. In *any* interaction we ask if the impact of one covariate depends on the level of another covariate. Typically, the two covariates are distinct (X_1 and X_2). Here, they are the same (X_1 and X_1) but the idea is the same. We are asking if the impact of X_1 depends on the level of X_1.

When working with curvilinear (or higher-order) relationships it's generally a good idea to center variables, and that's the practice we follow here. To assess the hypothesis that there is a curvilinear relationship between Dose and effect size we'll need two covariates

- Dose (C) is simply Dose centered at the mean (55.68).
- Dose $(C)^2$ is the square of Dose (C).

For information on how to create these variables see Appendix 5: Creating variables for interactions.

Figure 106 shows the model, Figure 107 shows the main results, and Figure 108 shows a plot of these results. While the program can compute the statistics for curvilinear relationships it cannot *plot* these relationships, and therefore Figure 108 was created in Excel™ (see appendix).

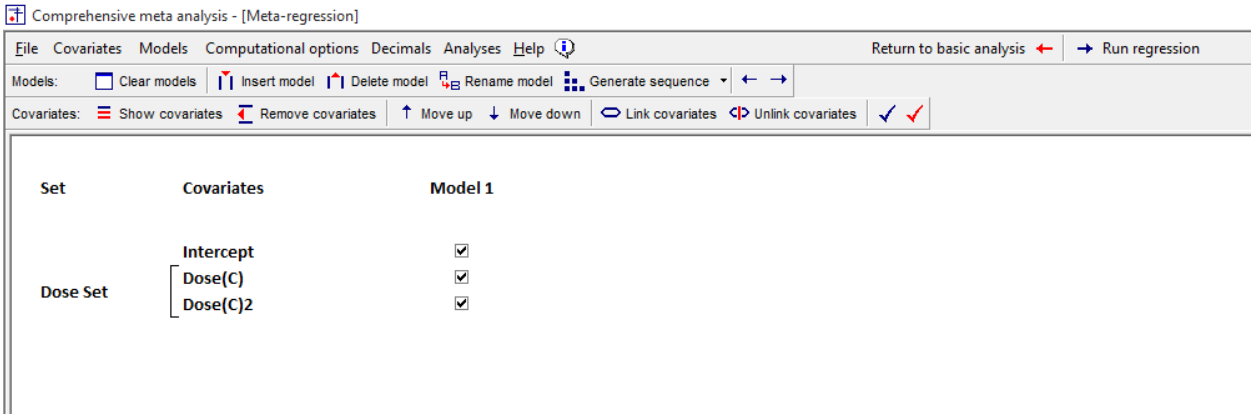


Figure 106 | Setup | Curvilinear relationship

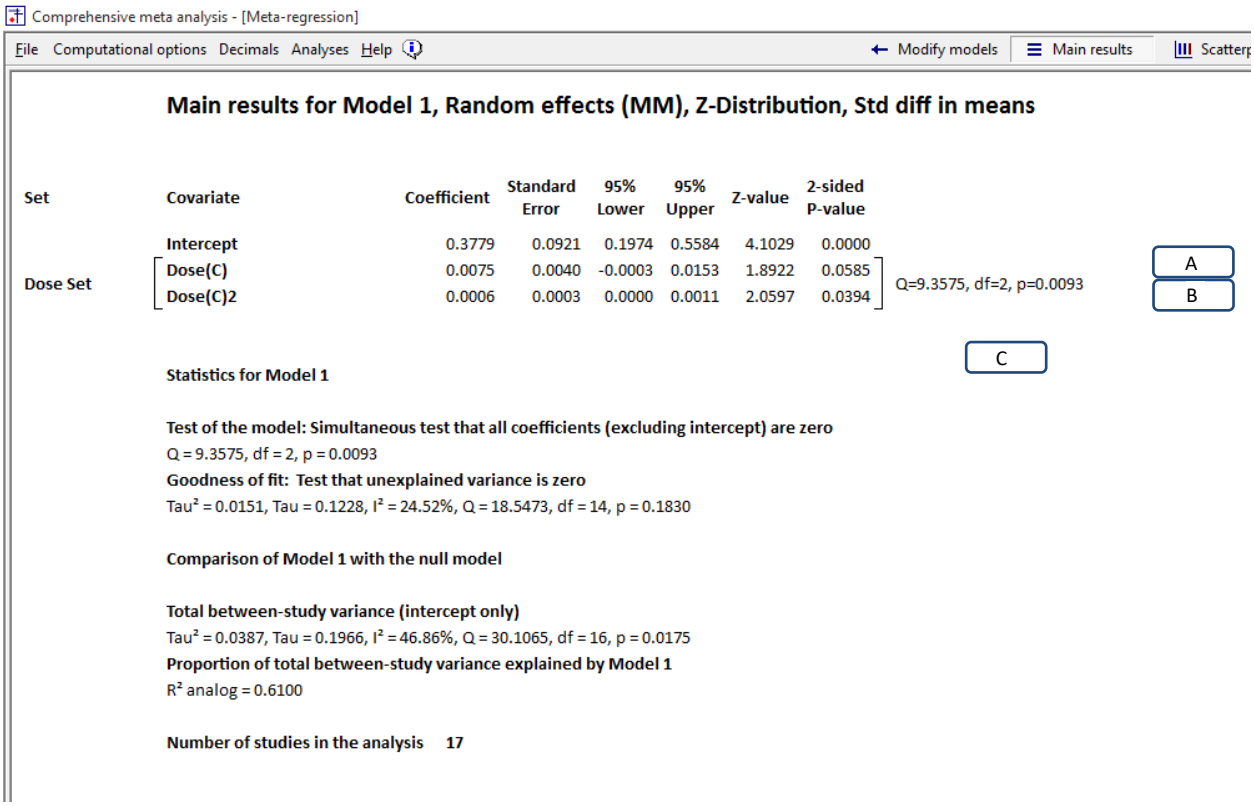


Figure 107 | Main results | Curvilinear relationship

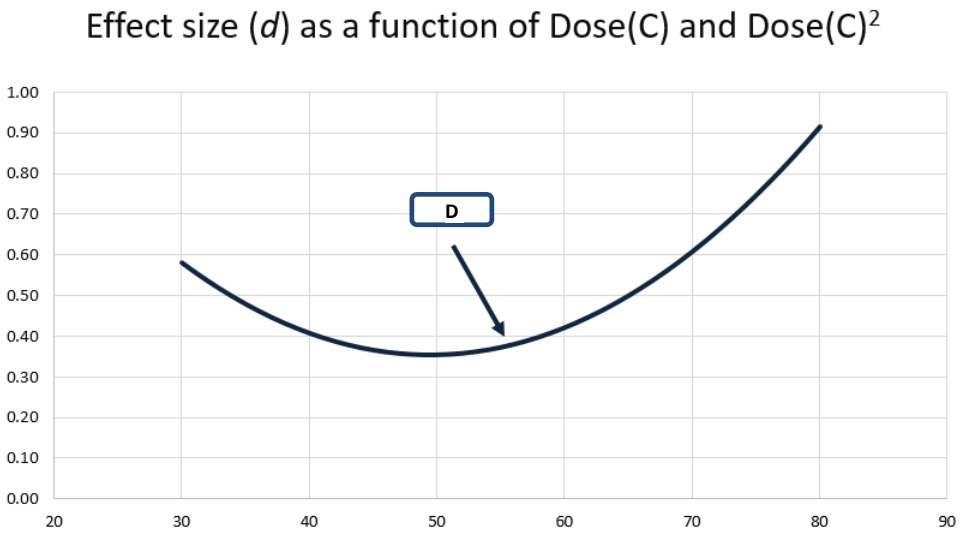


Figure 108 | Plot | Curvilinear relationship

We want to evaluate the linear component of this relationship, the curvilinear component, and then the two (as a set).

Dose-C

The line for *Dose-C* in Figure 107 [A] addresses the *linear* relationship between *Dose* and effect size. The *p*-value is 0.0585.

Note. Since we have included an interaction, the linear component varies as a function of *Dose*. Concretely, there are parts of the graph where the line has a downward slope and parts where it has an upward slope. The coefficient 0.0075 is the slope (or the tangent) at the point where *Dose-C* is zero and *Dose* is 55.86. This is indicated by an arrow [D] on the plot.

Dose-C^2

The line for *Dose-C*^{2} in Figure 107 [B] addresses the curvilinear component of the relationship between *Dose* and effect size. The line in the plot is curvilinear, with a slope that initially moves downward but then upward. Is this line a better fit for the data than a straight line would be? This is addressed by the *p*-value for this covariate, which is 0.0394.

Test of the set

The line for *Dose-C* addresses the linear component and the line for *Dose-C2* addresses the curvilinear component. We might also want to ask about the impact of the two together – Is dose (linear plus curvilinear) able to predict effect size. Since we're dealing with two covariates we use *Q* (rather than *Z*). The *Q*-value for the set is 9.3575 with *df* = 2 and *p* = 0.0093. We would reject the null that neither covariate is related to effect size and conclude that at least one is related to effect size.

Test of the model

The test of the model is the test that all coefficients are zero. In this case, since the only coefficients in the model are *Dose-C* and *Dose-C2*, the test of the model is identical to the test of the set. The *Q*-value for the set is 9.3575 with *df* = 2 and *p* = 0.0093. We would reject the null that neither covariate is related to effect size and conclude that at least one is related to effect size.

OBSERVATIONAL AND SMALL K (SUD)

COULD BE AN ARTIFACT – WHEN ADD SUD IT DISAPPEARS – EXAMPLE OF WHY DANGEROUS TO USE SMALL K. MAYBE MAKE THIS POINT IN THE SECTION ON MR BEING OBSERVATIONAL, AND NEEDING SUFFICIENT K

MISSING DATA

In regression for meta-analysis, as in regression for primary studies, there are many options for dealing with missing data. The program takes a very simple approach to missing data, as follows. If a study is missing data for the outcome or for any of the covariates in the covariate list that study is excluded from the analysis.

Note that this exclusion is based on *all* covariates listed on the main screen, *and not only on the covariates that are checked*. For example, consider the analysis shown here. The variables SUD and Dose have been added to the list of covariates. Therefore, any study that is missing data for SUD and/or Dose will be excluded from the analysis.

If a study is missing data for SUD and we want to include this study in the analysis, we *cannot* simply un-check SUD as in Figure 109. Rather, we need to highlight SUD and then click [Remove Covariates].

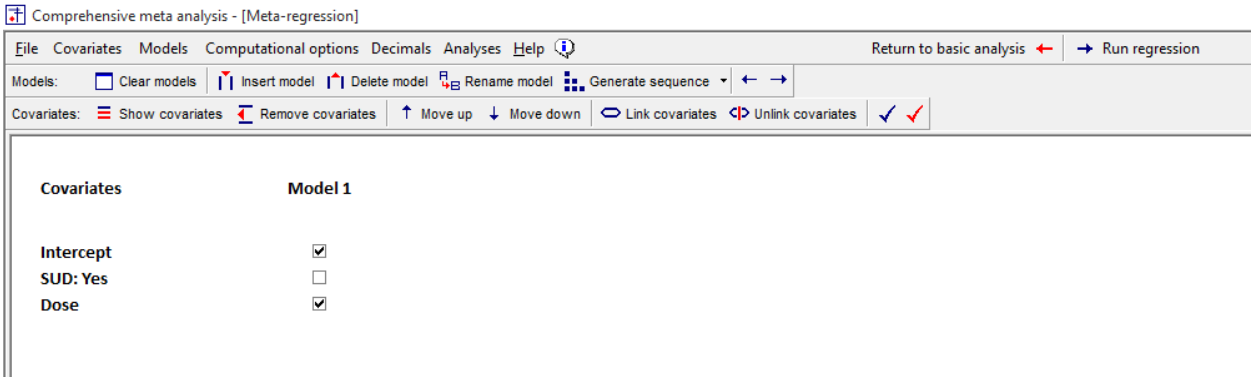
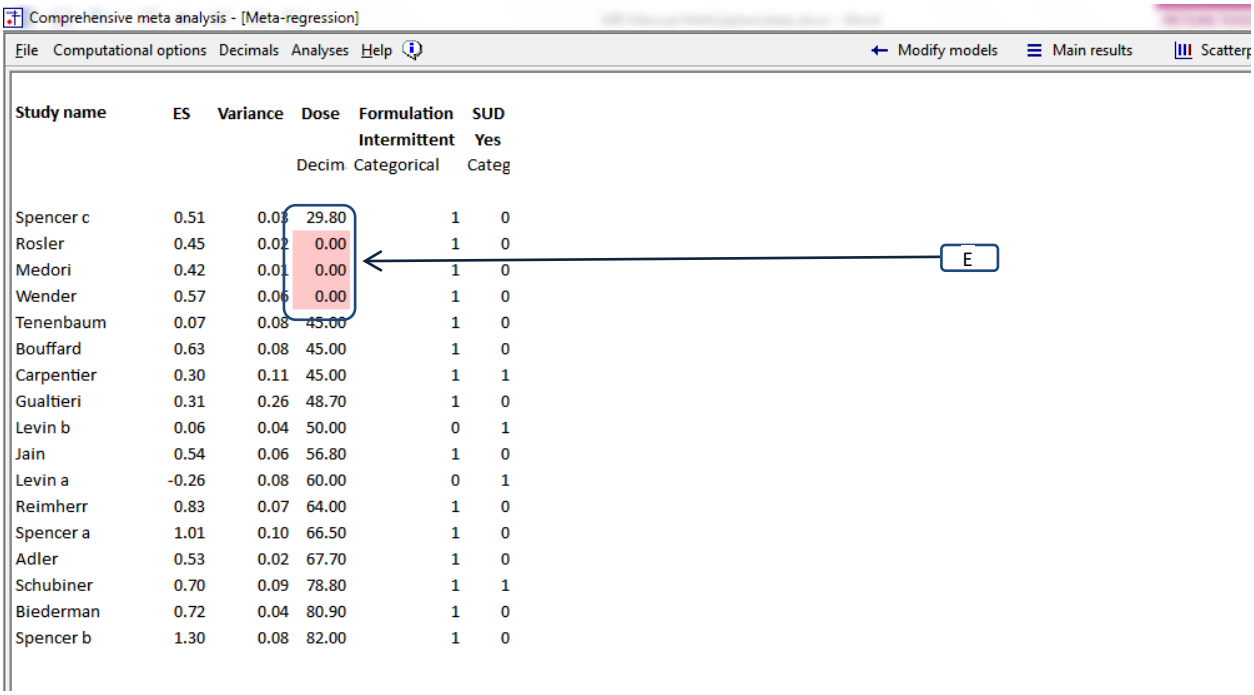


Figure 109 | Setup

- To see *which* studies have been excluded click More results > All data
- The program displays a line for every study in the database, and missing data points are highlighted in red (Figure 110).



Study name	ES	Variance	Dose	Formulation Intermittent	SUD Yes
			Decim	Categorical	Categ
Spencer c	0.51	0.03	29.80	1	0
Rosler	0.45	0.02	0.00	1	0
Medori	0.42	0.01	0.00	1	0
Wender	0.57	0.06	0.00	1	0
Tenenbaum	0.07	0.08	45.00	1	0
Bouffard	0.63	0.08	45.00	1	0
Carpentier	0.30	0.11	45.00	1	1
Gualtieri	0.31	0.26	48.70	1	0
Levin b	0.06	0.04	50.00	0	1
Jain	0.54	0.06	56.80	1	0
Levin a	-0.26	0.08	60.00	0	1
Reimherr	0.83	0.07	64.00	1	0
Spencer a	1.01	0.10	66.50	1	0
Adler	0.53	0.02	67.70	1	0
Schubiner	0.70	0.09	78.80	1	1
Biederman	0.72	0.04	80.90	1	0
Spencer b	1.30	0.08	82.00	1	0

Figure 110 | Table of missing data

This is a good way to identify the missing data and also to identify patterns of missing data.

- If data is missing primarily for one covariate across a lot of studies you may decide to remove that covariate from the analysis.
- If data is missing primarily for a few studies across many covariates you may decide to remove those studies and keep the covariates.

Of course, the decision to adopt one of these approaches or some other will depend on a host of factors, with attention paid to avoiding bias. However, the ability to identify the patterns of missing data is a crucial first step in this process.

There are more sophisticated methods for handling missing data, such as multiple imputation and selection models for non-ignorable missing data. While these are beyond the scope of this manual, these and other methods can be implemented using CMA. You would use an external program to determine the data value for each study, and then input this value via the data-entry screen.

In a more sophisticated version of this scheme you can create several variables based on the same variable, but using different approaches to missing data. For example, suppose the initial variable is *Dose*. You can create one variable called *DoseA* that replaces missing data with the mean, and another variable called *DoseB* that replaces missing data with another imputed score. Then, in any given model you would use one or the other, but not both.

PART 16: FILTER STUDIES

MAYBE MAKE A SECTION OF THE BOOK LABELED – USING CMA

In some cases you may want to run a regression using a subset of the data. For example, you might want to limit the analysis to studies that were performed within the past ten years, or to studies that employed specific variants of the intervention or that enrolled persons from specific populations.

This procedure is called “Filtering”, in that we create a filter, and only studies that pass through the filter are submitted for the regression. The process is actually very simple, but requires that you understand the relationship among three distinct modules in the program. This is shown schematically in the following three figures.

1. In the data-entry module we enter data for all studies (Figure 111)
2. In the main analysis module we can create filters (Figure 112)
3. Studies that pass through the filters are submitted to the regression module (Figure 113)

Data-Entry

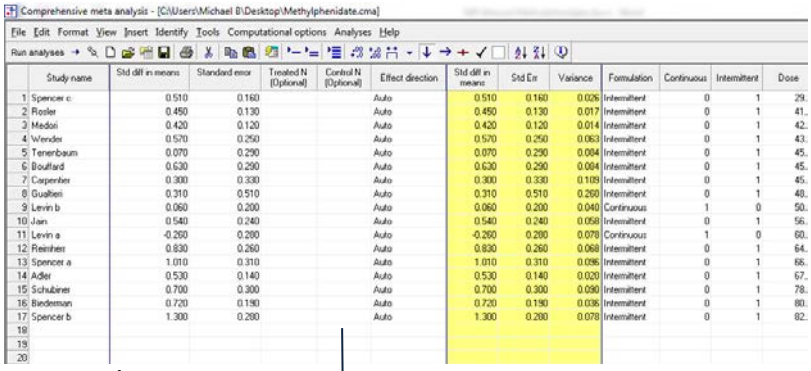


Figure 111 | Data entry

Basic-Analysis
All filtering is done here

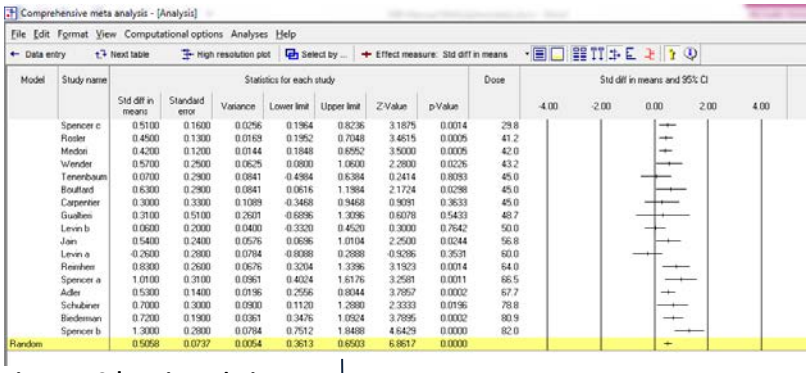


Figure 112 | Basic analysis

Meta-regression

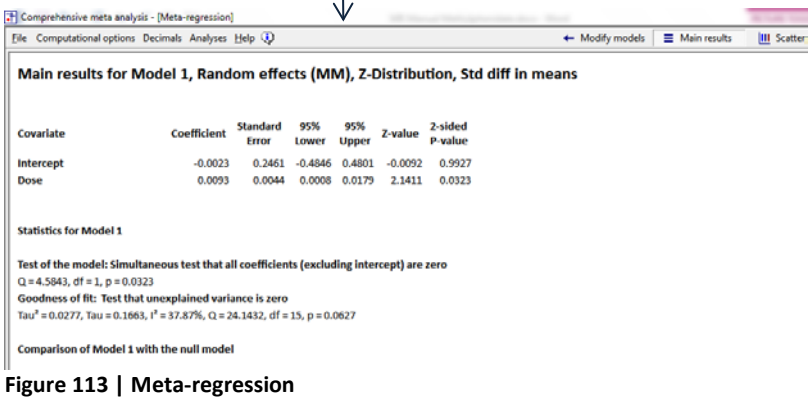


Figure 113 | Meta-regression

We provide a few examples of filtering

Example 1

Suppose you want to exclude two specific studies (Levin a and Levin b) by name.

On the main analysis screen (Figure 114), click Computational options > Select by > Study name”

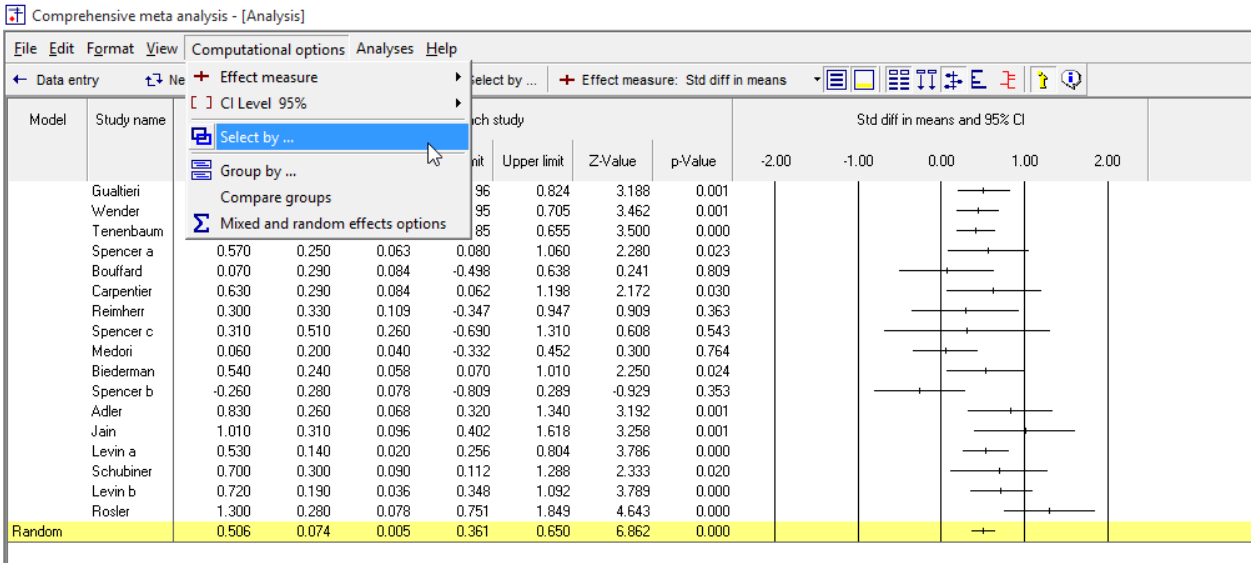


Figure 114 | Select by study name

In Figure 115, select or de-select studies and click [Ok]

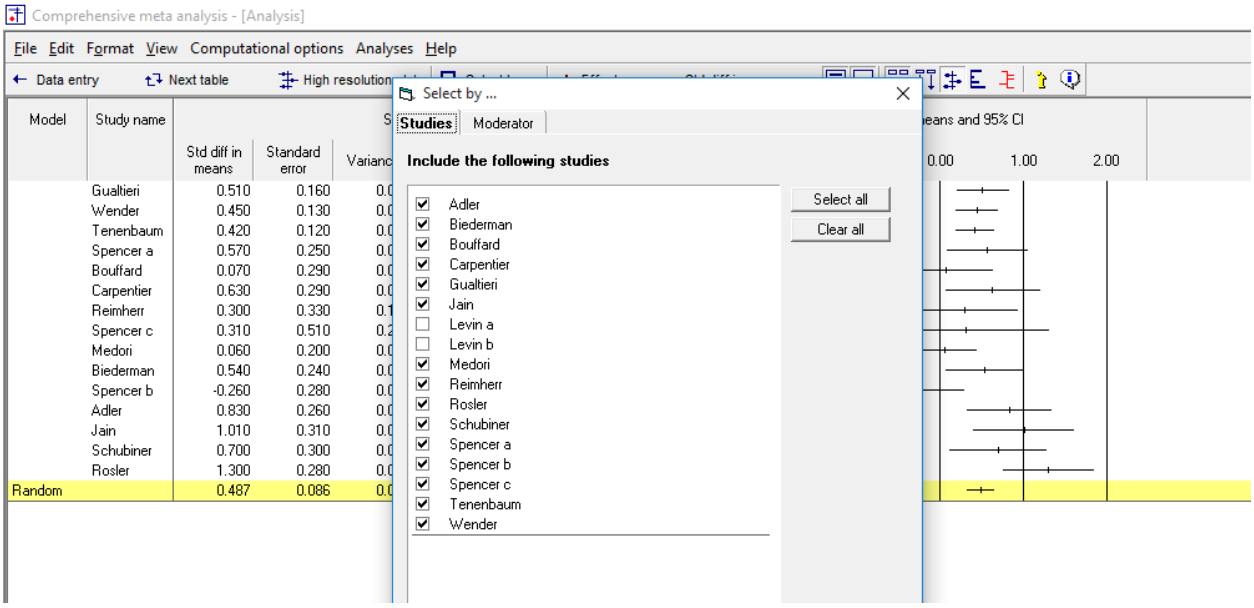


Figure 115

The analysis is now based on the remaining studies. When you proceed to meta-regression, only these studies will be transmitted.

Example 2

You can filter studies based on existing moderators. For example, suppose you wanted to run an analysis using studies that excludes studies where the formulation was continuous.

- Click Computational options > Select by >
- Select Moderator > Formulation
- De-select [Continuous]

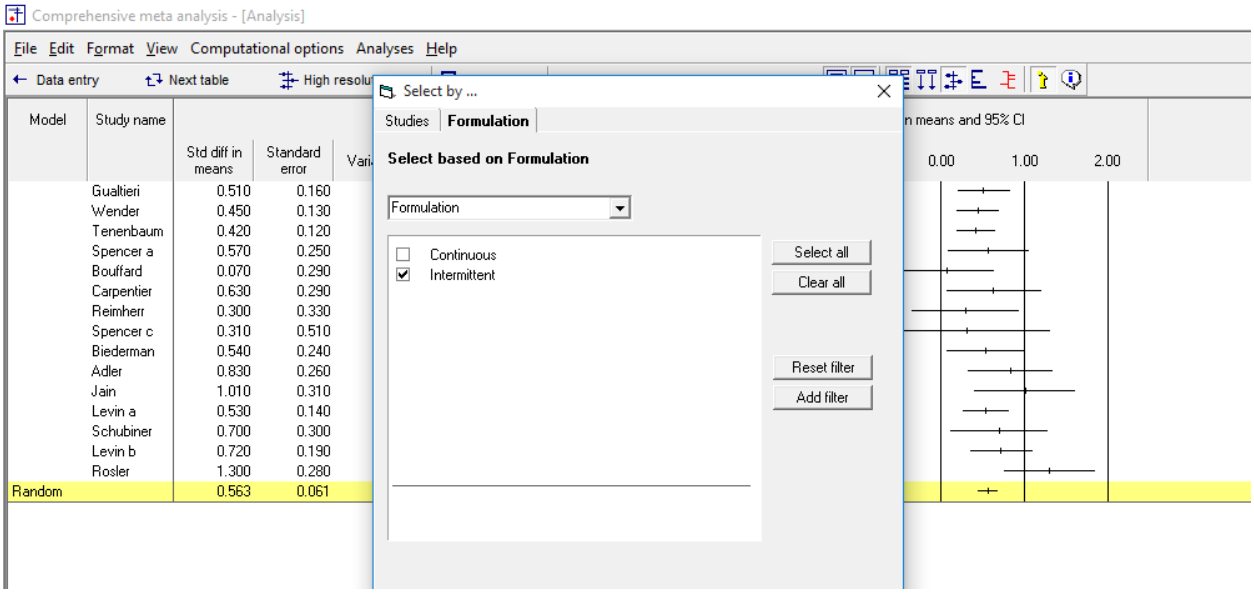


Figure 116 | Select by moderator

Note that you can add tabs with additional moderators. A study will only be included if it meets the criterion on all tabs.

Example 3

Suppose that you want to run a series of analyses using a specific subset of the studies which are not identified by any existing moderator. A simple solution is to create a new moderator specifically for this purpose.

Create a categorical moderator (Set-A) and code each study as belonging to this set (or not). Create a second categorical moderator (Set-B) and code each study as belonging to this set (or not). Then, on the main analysis screen, it's a simple matter to switch back and forth among these sets.

Keep in mind that the sets are cumulative. If you're working with Set-A and then want to move on to Set-B, either remove Set-A from the selection wizard or ensure that all check-boxes for Set-A are ticked.

DEFINING SEVERAL MODELS

MOVE

Typically you will create one prediction model, which is the list of covariates to be included in the analysis. Then, you might try another model, and another, working one model at a time.
The program offers another option – to define a number of prediction models at once, and then run them all simultaneously. To create additional models click [Insert Model] on the [Models] toolbar [A]. Then add check-marks to indicate which variables will be included in each model.

Here, for example, the user has defined one model that includes only the intercept [B], a second that adds Dose [C], a third that adds Duration [D] and a fourth that adds the Interaction [E].

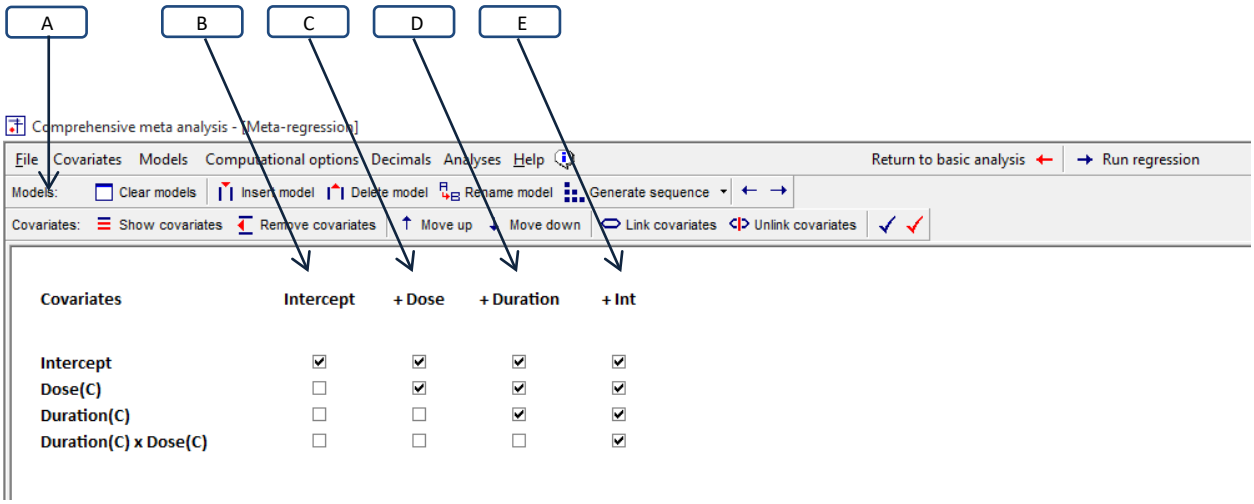


Figure 117 | Defining several models | Setup

In this example we've named the first model "Intercept" since this is the only covariate. The second model is named "+ Dose" since this includes intercept plus Dose, and so on. Note that the name is simply intended as a mnemonic. We could have called the model by any names.

When you run the analysis, the program creates a tab corresponding to each model as shown in the following frames.

In Figure 118 the user has clicked on the tab [B] for the model labeled “Intercept”. The program displays the analysis for the first model.

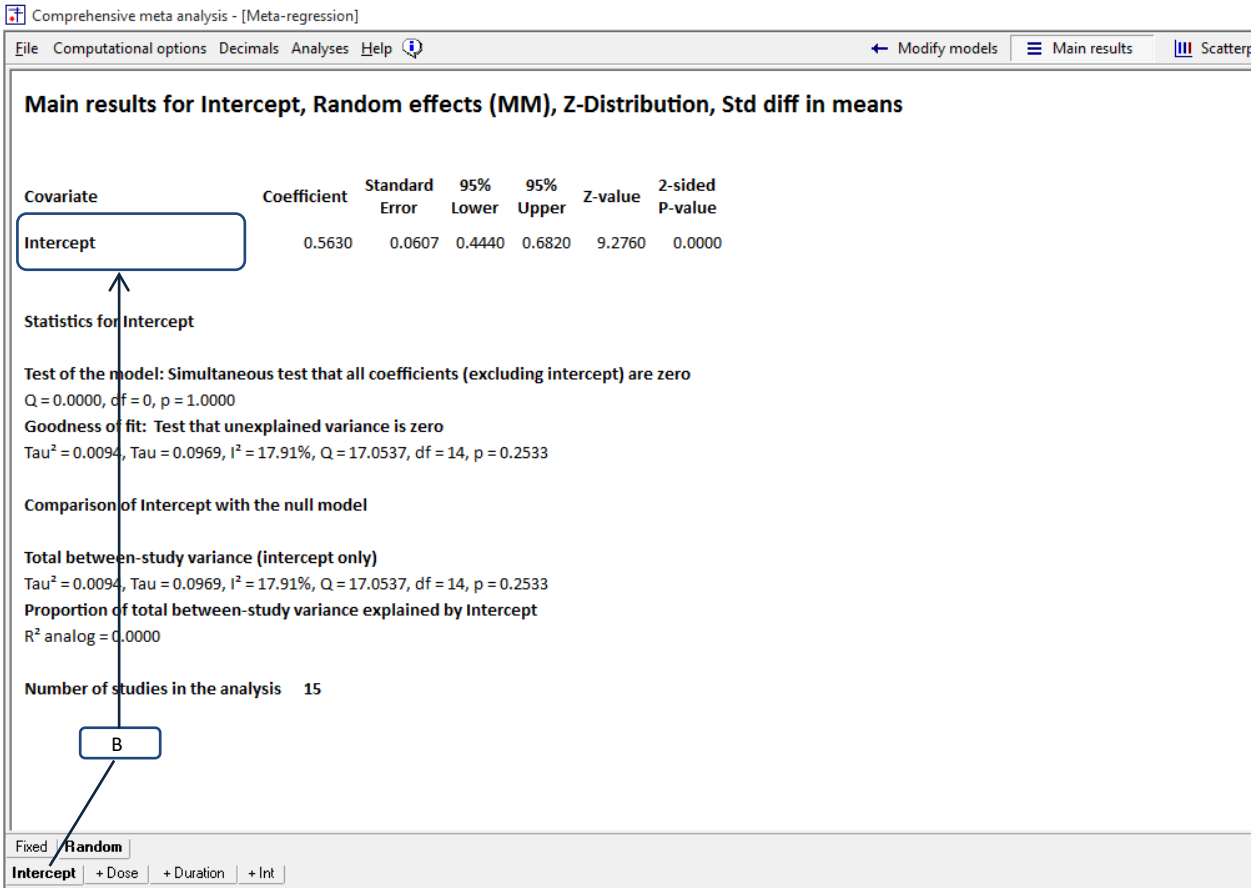


Figure 118 | Defining several models | Main-analysis | Intercept

In Figure 119 the user has clicked on the tab [C] for the model labeled "+ Dose". The program displays the analysis for the second model.

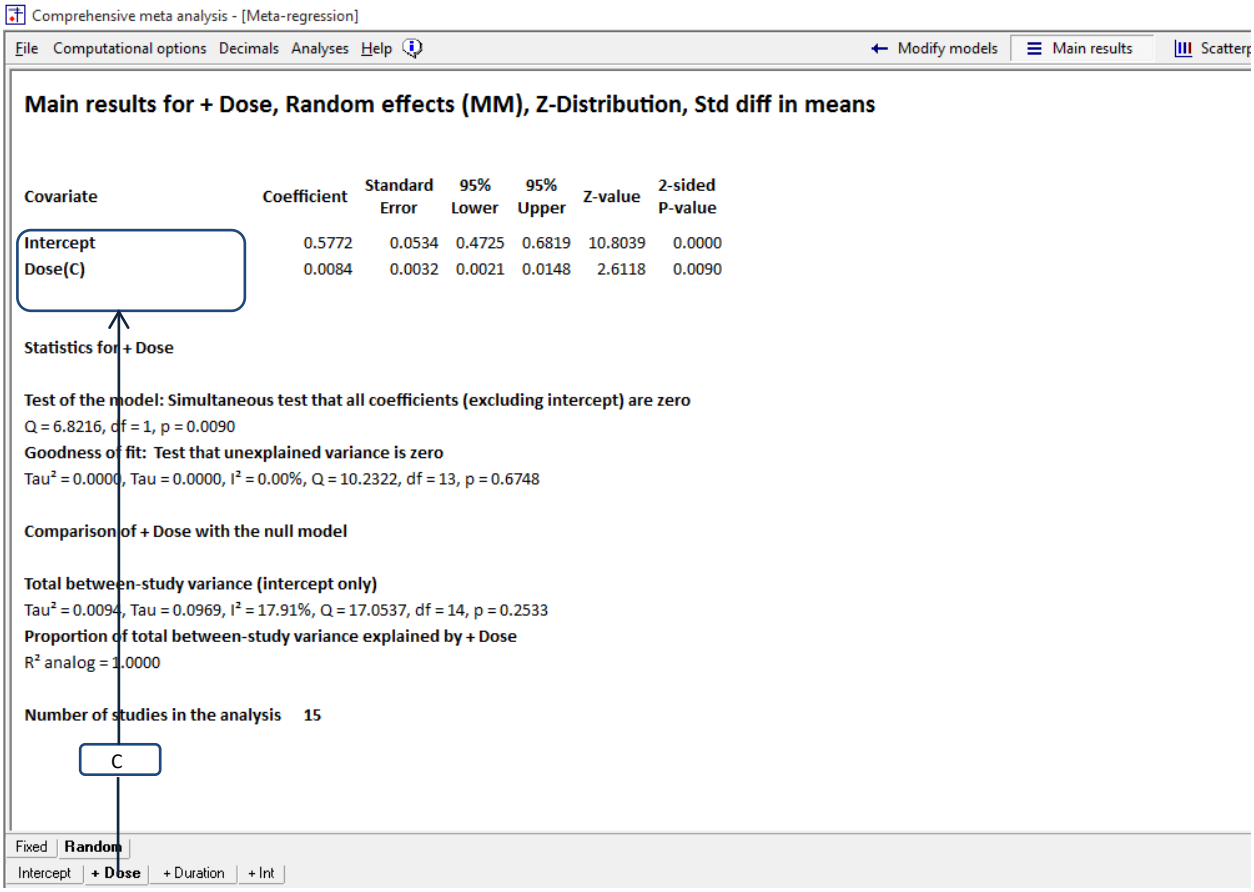


Figure 119 | Defining several models | Main-analysis | Intercept + year

In Figure 120 the user has clicked on the tab [D] for the model labeled "+ Duration". The program displays the analysis for the third model.

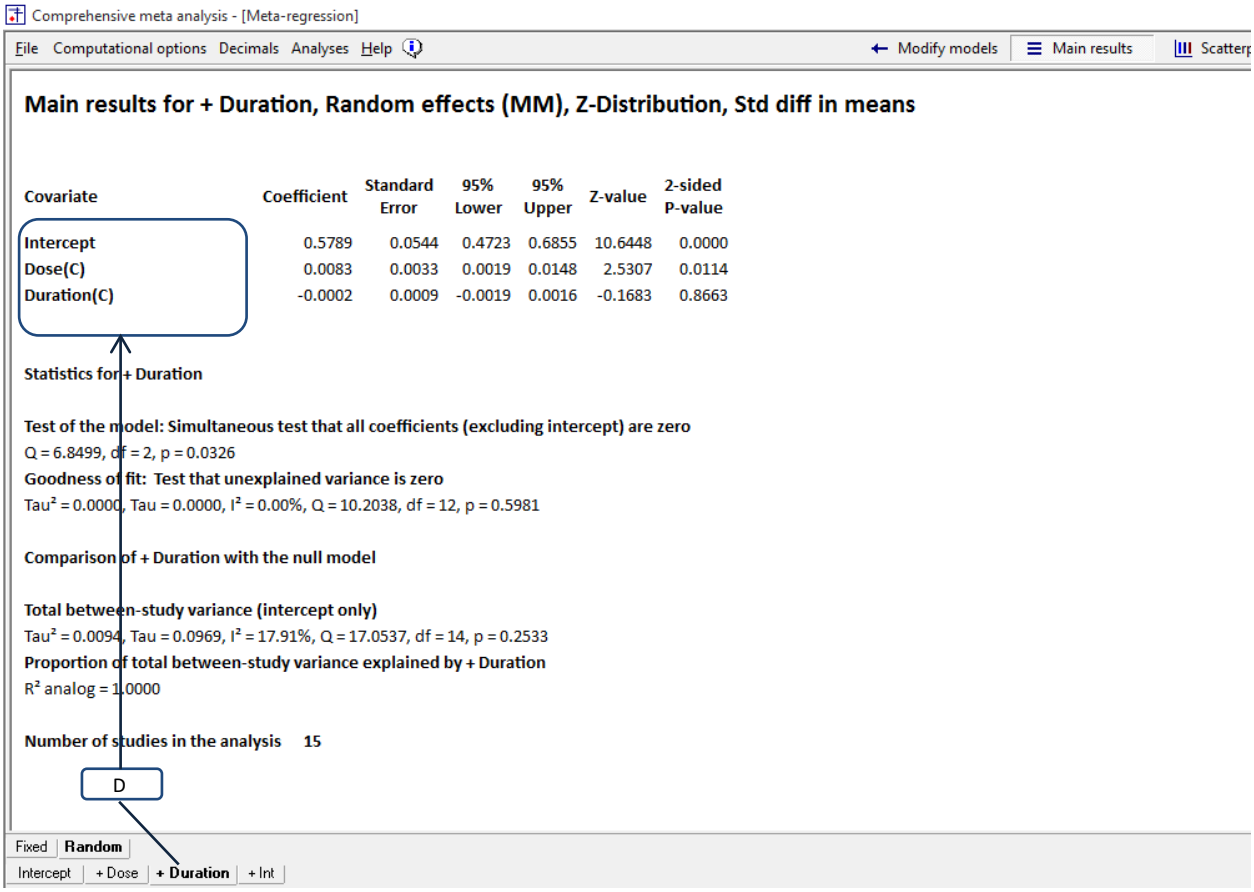


Figure 120 | Defining several models | Main-analysis | Intercept + year + dose

In Figure 121 the user has clicked on the tab [E] for the model labeled "+ Int". The program displays the analysis for the fourth model.

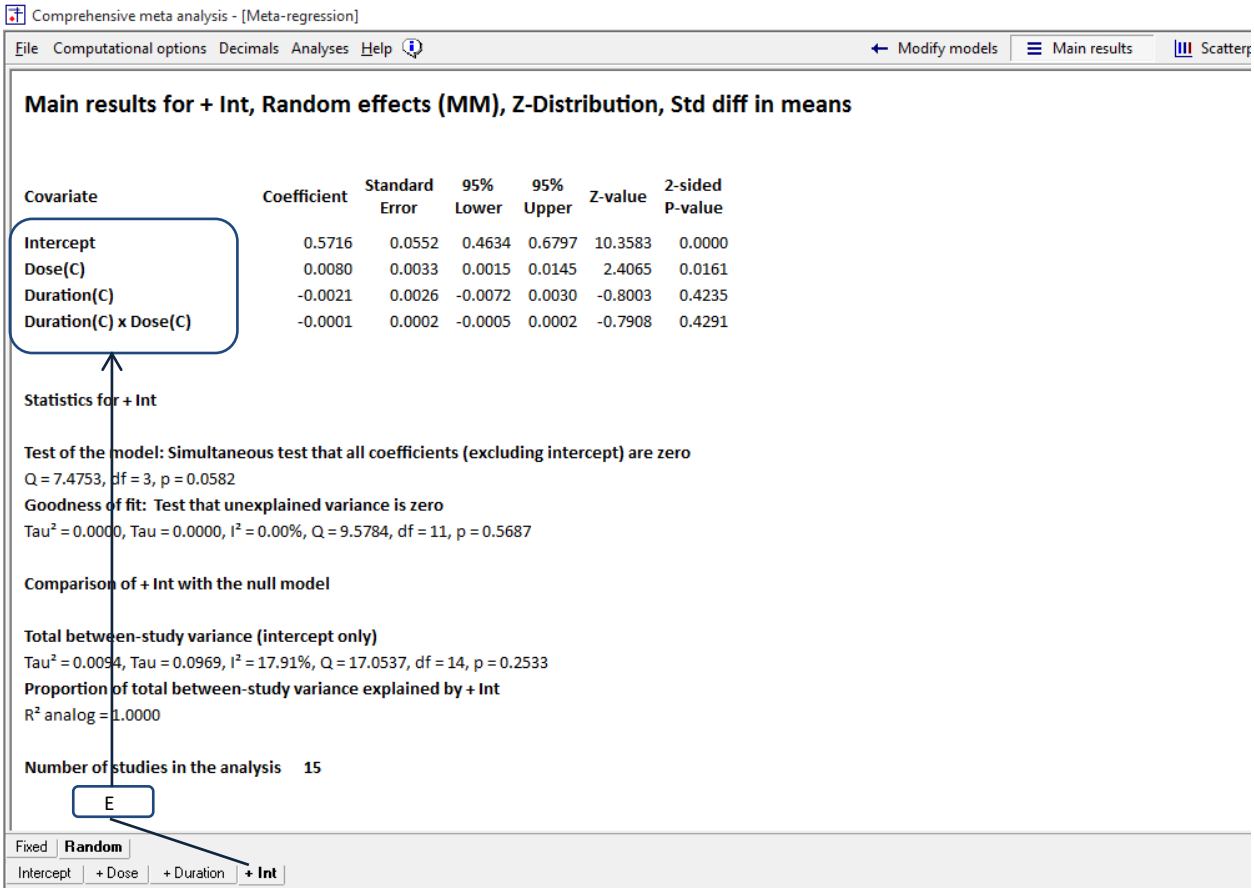


Figure 121 | Defining several models | Main-analysis | Intercept + year + dose

Similarly, compare Figure 122 where the user has selected the tab for +Dose, with Figure 123 where she has selected the tab for +Duration. In both cases we've plotted effect size on Dose but in the first case the equation is based on Dose alone, whereas in the second it's based on Dose controlling for Duration. In this example we see the impact of Dose as evidenced by the plot and the equation is essentially unchanged by the addition of Duration.

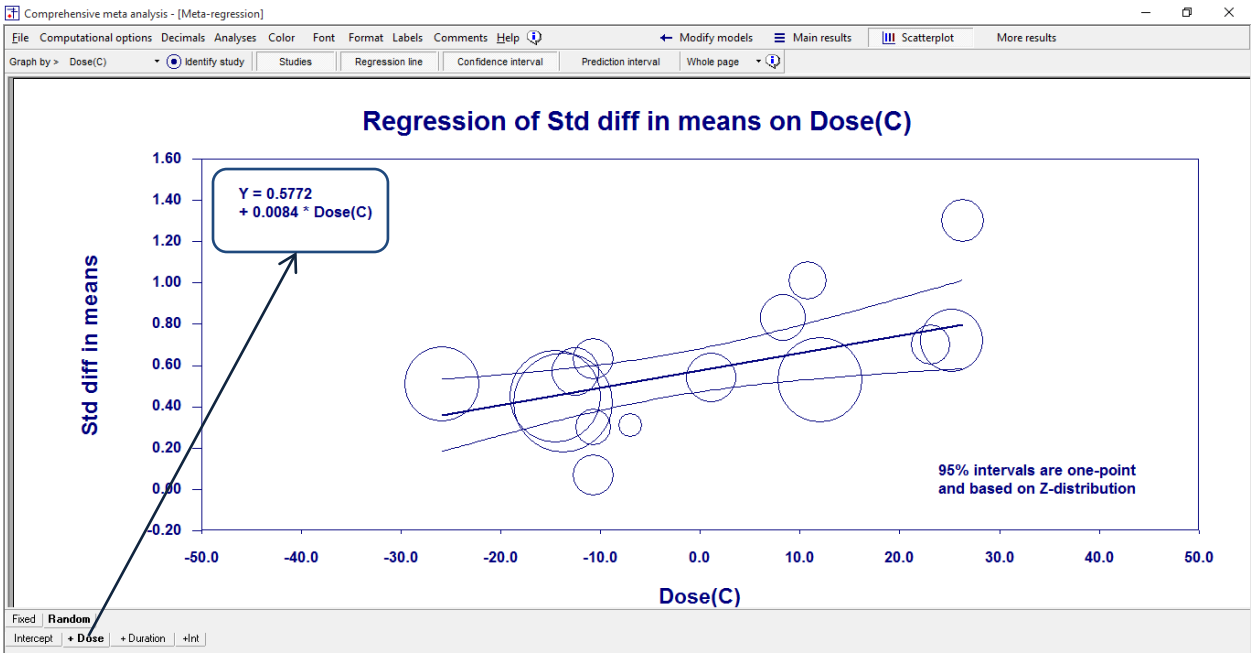


Figure 122

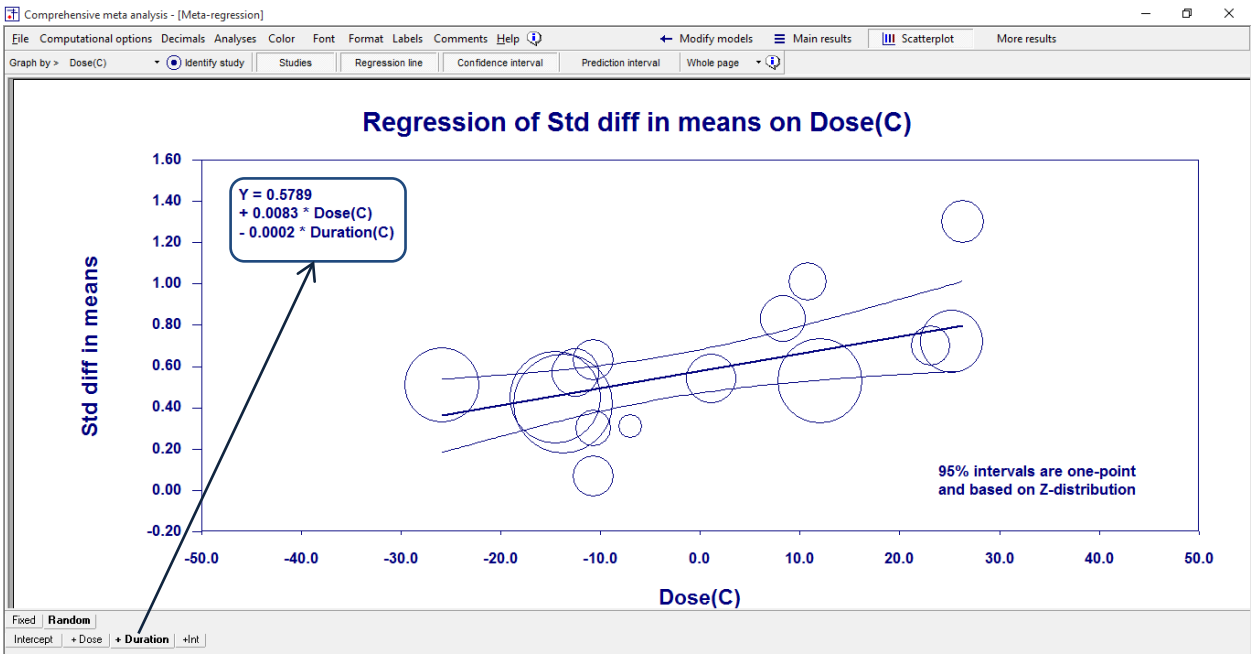


Figure 123

Why would we want to define more than one model?

In the running example we defined four models on the main screen. Why take this approach, rather than simply working with one prediction model at a time? There are two reasons why this option may be useful.

First, the program summarizes the results of all models on one screen (Figure 124). To navigate to this screen click More results > [Summary]

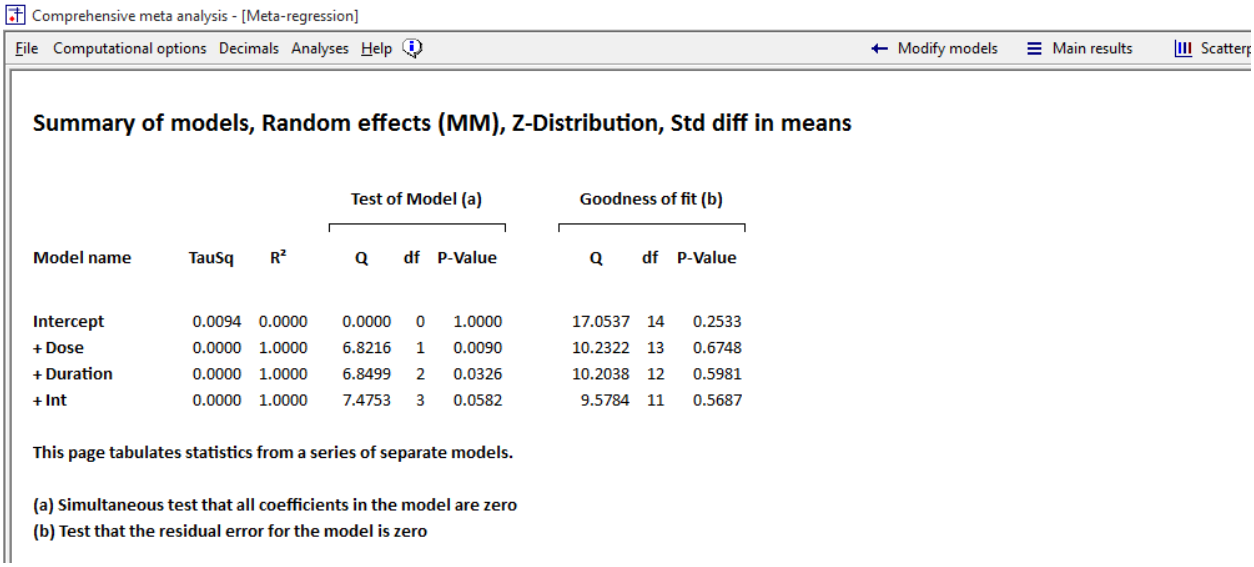


Figure 124 | Defining several models | Main-analysis | Intercept + year + dose

HERE, WE SEE …

Second, the program displays a test for the difference in the explanatory power of the models (Figure 125). For example, the cell indicated by [F] compares the model that includes intercept and dose with the one that adds Duration as well. This option is only available when one model is a subset of the other. When this is not the case, the corresponding cell will be left empty.

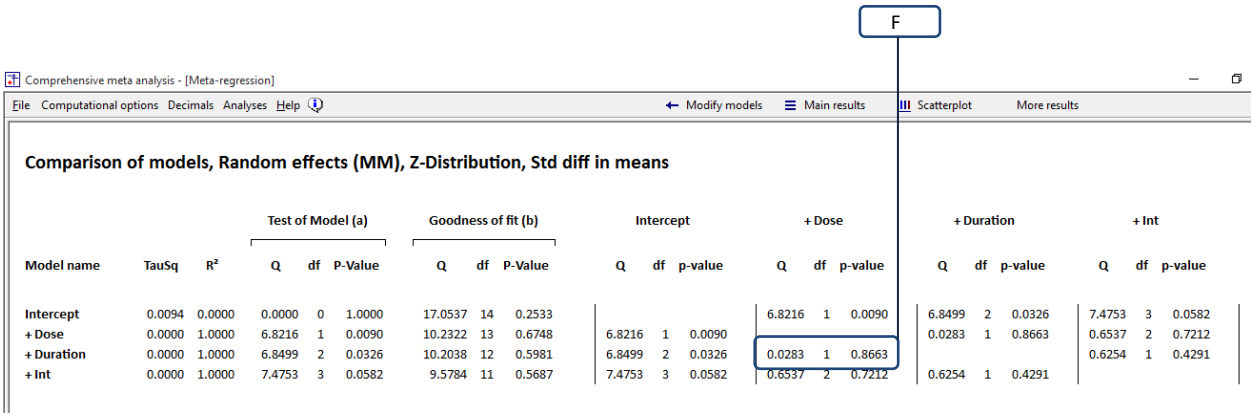


Figure 125 | Defining several models | Main-analysis | Intercept + year + dose

How do we choose what covariates to include in each model? This depends on the questions we want to address.

For example, suppose the primary goal of the analysis is to assess the impact of treatment.

- One series of covariates such as mean age and location is seen as noise
- One series of covariates represents treatment condition
- One series (such as dose by treatment) represents potential interactions

We might define one model as “Nuisance”, a second as “Plus Treatment”, and a third as “Plus interactions”. Then, the summary screen provides a quick look at the three models while the comparison screen shows the statistical tests of the differences among them.

CREATING A SERIES OF MODELS AUTOMATICALLY

Earlier, we showed that the user can Insert and define additional models manually. Alternatively, the program can create a series of models automatically to match two common sequences. You would simply move the variables onto the main screen and then choose one of two options.

Incremental sequence

The incremental sequence is Figure 126. The first model includes the intercept only, and then one covariate is added at each step. To create this sequence click Generate sequence > Incremental sequence [G]. The name for each variable is "plus" the name of the new covariate. Use [Rename model] to rename the models.

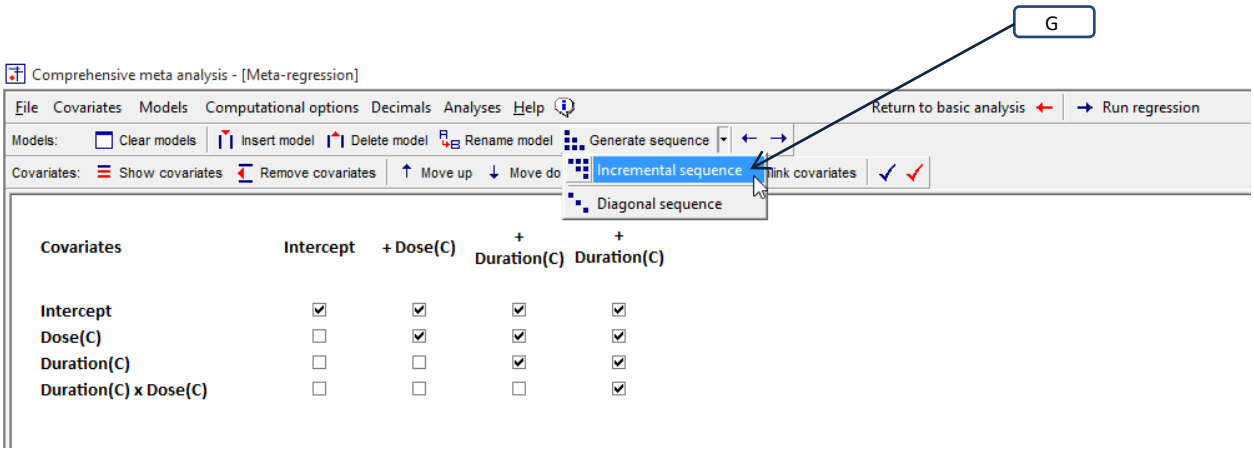


Figure 126 | Defining several models | Main-analysis | Intercept + year + dose

Diagonal sequence

The diagonal sequence is Figure 127. Each model includes the intercept plus one covariate. To create this sequence click Generate sequence > Diagonal sequence [H]. The name for each variable is that variable alone. Use [Rename model] to rename the models.

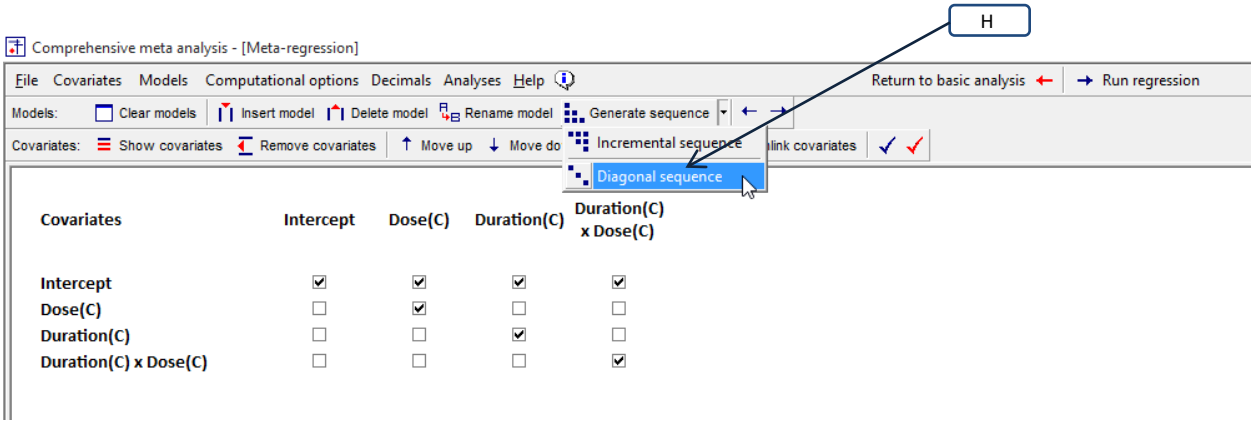


Figure 127 | Defining several models | Setup| Year or Dose

Note.

When you run multiple models, always be sure to select the desired tab at the bottom when studying the results.

- Screens that present results for one predictive model will change as the user selects one or another model using the tabs at bottom.

- Screens that collate results for all one predictive models do not change as the user selects one or another model using the tabs at bottom.

Table 2

Screen	Varies by model	Identical for all models
Main results	X	
Scatterplot	X	
R2 graphic	X	
Covariance	X	
Correlation	X	
Diagnostics	X	
All studies		X
Valid studies		X
Increments	X	
Models summary		X
Compare models (detailed)		X
Compare models (p-values)		X

PART 19: SOME CAVEATS

COVARIATES AT THE LEVEL OF THE INDIVIDUAL

In a meta-regression, the covariate is defined at the study level. For example, the Dose for a study might be entered as 50 mg.

If the Dose is actually set at 50 mg for all patients in the study, then this works well. On the other hand, if the dose varied from patient to patient, and 50 mg is the mean dose, then by entering the mean we are losing some potentially important information.

Similarly, analyses that use gender as a covariate may enter a number that reflects the proportion of patients who are female. Again, if gender is an important predictor, then we lose a lot of information in this way.

Sometimes, it's possible to bypass this problem. In the gender example, if a study reports the effect size separately for males and females we could enter the data separately for each subgroup, with a moderator coded either Male or Female. When this is not possible, then some information will be lost.

STATISTICAL POWER FOR META-REGRESSION

Statistical power is the likelihood that a test of significance will reject the null hypothesis. In the context of a meta-regression this would refer to the likelihood that the *Z*-test of a single covariate or the *Q*-test of a set of covariates will yield a statistically significant *p*-value. Power is driven by the magnitude of the effect and the precision of the estimate.

While there may be a perception that statistical power for a meta-analysis is consistently high, this perception is incorrect. The perception is driven by the fact that a meta-analysis may include a large number of subjects, which tends to yield a very precise estimate of the effect size. However, the perception is incorrect because precision is also driven in part by the number of studies, and this number may not be large.

In the case of a meta-regression power may also be hurt by the number of covariates. To have good power we need not only a sufficient number of studies, but also a sufficient number of studies relative to the number of covariates.

The key point is that we cannot assume that power for a meta-regression is good. Therefore, if the impact of a covariate (or a set of covariates) is not statistically significant, we should take the same approach as we would take in a primary study. That is, we would conclude that the null hypothesis may or may not be true.

Hedges and Pigott, 2004, 2001

ECOLOGICAL FALLACY
ALSO, SEE HETEROGENEITY PAPER (CHAPTER)

MULTIPLE COMPARISONS

Consider the case where we run a meta-regression with several covariates, and test each using a criterion alpha of 0.05. The type-1 error rate for any single comparison is 0.05 but (if the null hypothesis is true) the type-1 error rate across *all* comparisons will be greater than 0.05.

This issue has been discussed extensively for multiple regression in primary studies, and the approaches that have been adopted there can be applied for meta-regression as well. Here, we list some of the common approaches. This list is not exhaustive and we provide only an outline of the approaches.

- Do not make any formal adjustment but apply some common-sense. For example, one significant *p*-value in forty tests would be suspect.

- Start with a test that asks if *any* of the relevant coefficients is non-zero. This could be test of the full model, or of the covariates that represent the treatment effect (as a set). If this test is statistically significant then proceed to look at individual covariates.

- Use a stricter criterion for significance (for example, a criterion alpha of 0.01 rather than 0.05 for five tests). Variations on this theme include the Bonferroni approach and the Schefee approach.

Others (e.g., Rothman 1990) suggest that there are many cases where we can safely ignore the problem of multiple comparisons.

Hedges and Olkin, 1985

PART 20: TECHNICAL APPENDIX

ADD BCG EXAMPLE

SHOW HOW TO PLOT RATIOS IN EXCEL

APPENDIX 1: THE DATASET

The motivating example in this book is the ADHD dataset.

- The Excel™ version is called ADHD.xls
- The CMA version is called ADHD.cma

The original analysis is described in _____. The authors located 19 studies that met their criteria, omitted one for lack of useable data, and performed their analysis on 18 studies. We omitted one additional study (Kuperman) because this study did not report the dose. This allowed us to base all analyses on the same set of 17 studies.

The original dataset includes only a few moderators, but we created additional moderators for purposes of this book. For example, the original dataset includes Dose but we created a variable that is a centered version of Dose, a variable that groups studies into Low-Dose vs. High-Dose, and so on.

Table 3

Variable	Type	Description
Variables related to dose		
Dose	Decimal	Dose of methylphenidate
Dose(C)	Decimal	Dose, centered at the mean dose
Dose(C) sq	Decimal	The square of Dose(C)
High	Categorical	"Low" (29 to 51) or "High" (56 to 82)
High (I)	Integer	"0" (29 to 51) or "1" (56 to 82)
Range	Categorical	"Low" () "Moderate" () or "High" ()
Variables related to duration		
Duration	Integer	Study duration in Days
Duration (C)	Integer	Days, centered at the mean
Long	Categorical	"Long" (5 to 36) or "Short" (42 to 196)
Long (I)	Integer	0 (5 to 36) or 1 (42 to 196)
Variables related to formulation		
Formulation		Continuous or intermittent dosing
Variables related to SUD		
SUD	Categorical	Included substance-abuse patients
SUD(I)	Integer	0 if SUD is "No". 1 if SUD is "Yes".
Interactions		
High(I) x Long(I)	Integer	High(I) x Long(I)
High(I) x Days(C)	Decimal	High(I) x Days(C)
Duration (C) x Dose(C)	Decimal	Duration (C) x Dose(C)

Notes

(C) Variable is centered at the mean
(I) Defined as an integer variable

Types of variables

Categorical	Studies belong to discrete groups, such as "N" and "Y". If this variable is entered into the analysis the program will create automatically covert this into a dummy-variable
Integer	Studies have values on an integer scale, and can take on any whole number.
Decimal	Studies have values on a continuous scale and can take on any whole or decimal number.

APPENDIX 2: UNDERSTANDING *Q*

REPLACE THIS WITH TEXT FROM CHAPTER ON HETEROGENEITY

Introduction

In a primary study we compute the sum of squares (*SS*) using

$$SS = \sum (X - Y)^2. \tag{1.82}$$

That is, we compute the deviation of each study from the mean, square the deviation, and then sum these squares across subjects. Once we have the SS we use it to derive various indices that reflect specific aspects of the variation. Specifically, we compute the variance (S2) and the standard deviation (S) using

$$S^2 = \frac{SS}{df} \tag{1.83}$$

and

$$S = \sqrt{S^2} \ . \tag{1.84}$$

In a meta-analysis we take the same approach, except that we want to assign more weight to the more precise studies. To do this, we divide each deviation by that study's standard error to get a *Z* value. Then, we square the *Z* values and sum these values across studies to get the *weighted* sum of squares, or *Q*. That is,

$$Q = WSS = \sum \left(\frac{X - Y}{SE_X} \right)^2 \tag{1.85}$$

Once we have computed *Q*, we can use it to derive a number of indices (such as the *p*-value, I^2, and T^2) that reflect specific aspects of heterogeneity.

First, we can test the *Q* value for statistical significance. Under the null (that all studies share a common true effect size, and that all the observed variance is due to sampling error) the expected value of *Q* is equal to the degrees of freedom, and *Q* follows the chi-square distribution. We can use *Q* and df to obtain a p-value. If the p-value is less than 0.10 we conclude that the true effect size probably varies across studies. (But see the discussion of fixed-effect and random-effects models in the text).

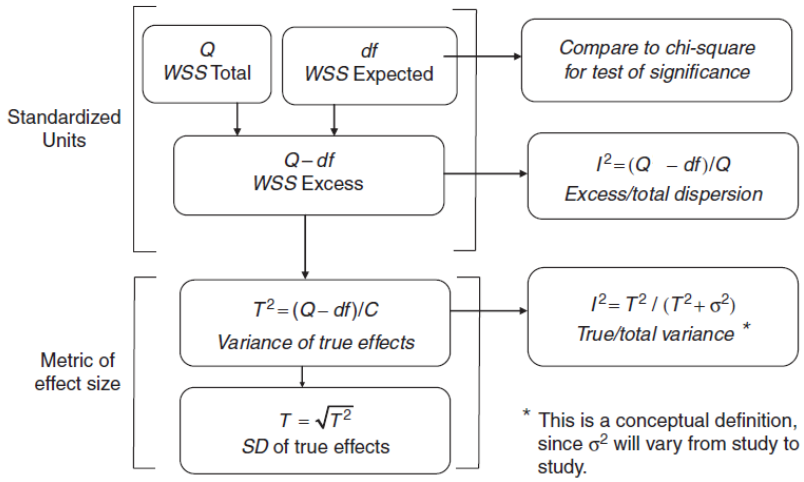


Figure 16.3 Flowchart showing how T^2 and I^2 are derived from Q and df.

Figure 128 | Flowchart showing how T^2 and I^2 are derived from Q

Second, we can use Q to estimate the amount of between-study variance. The value of Q we would expect to see based on sampling error is simply the degrees of freedom. It follows that if the observed value of Q exceeds the df, the excess can be attributed to between-study variance. This value ($Q - df$), is the basis for the following.

Variance and standard deviation of true effects

To compute T^2, the between-studies variance, we start with ($Q - df$). We multiply this difference by a factor (called C) that puts it into the metric of the effect size, squared. If T^2 is the variance of effect sizes between studies then its square root, T, is the standard deviation of effect sizes between studies.

Ratio of observed variance to true variance

To compute I^2, the ratio of true to total variance, we start with ($Q - df$). If Q reflects the total WSS and ($Q - df$) reflects the WSS between studies, then the ratio $(Q - df)/Q$ is by definition the ratio of true/total variance, called I^2. By convention we multiple I^2 by 100 and express it as a percentage (0% to 100%).

Proportion of true variance explained by the predictors

In primary studies, R^2 is the proportion of variance explained by predictors. In meta-regression the analog to R^2 is the proportion of true variance explained by predictors, based on the explained variance as a proportion of the original variance.

Here, we show how we compute Q and then use it to derive other statistics. We do this for (a) a simple analysis, (b) a subgroup analysis, and (c) a regression, to highlight the fact that the basic computation is the same in all three cases.

Case A | A simple analysis

To estimate the mean effect across all studies we run the regression with only the intercept. On the tab for fixed-effect weights [A] the intercept is shown as 0.4928 [B]. This is the mean effect and also (by definition) the predicted effect size for all studies. Note that we use the fixed-effect analysis to compute *Q* even if we will be using random-effects weights in the subsequent analysis.

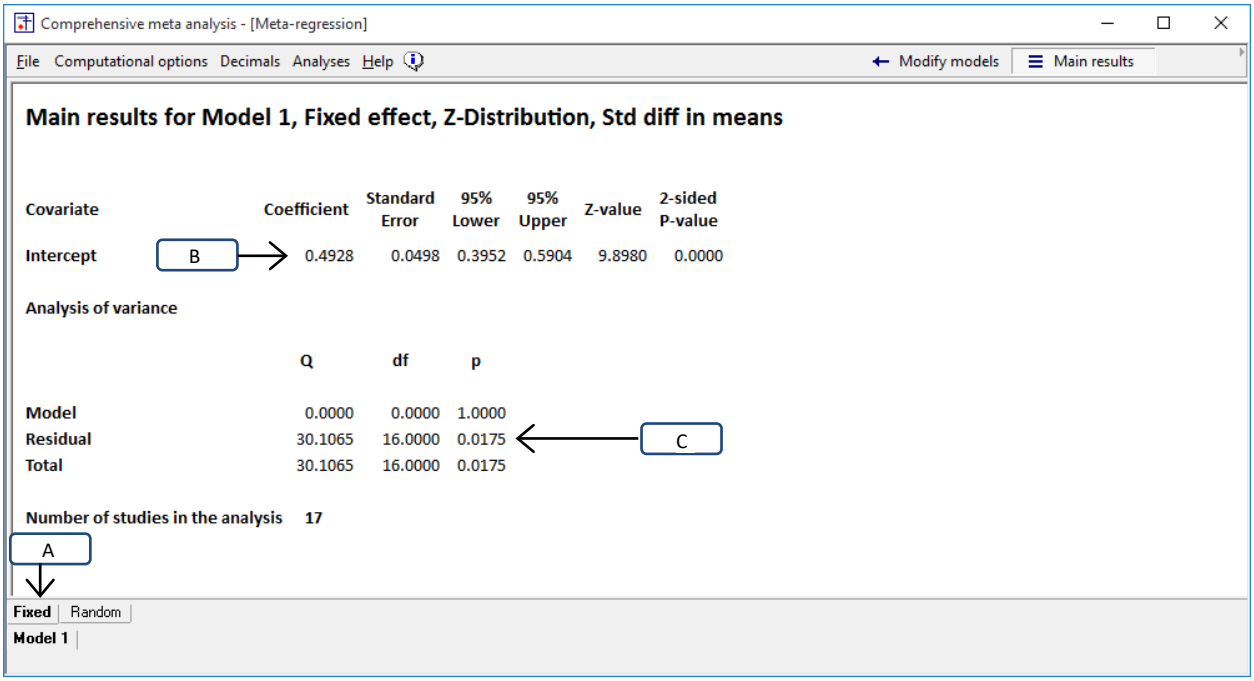
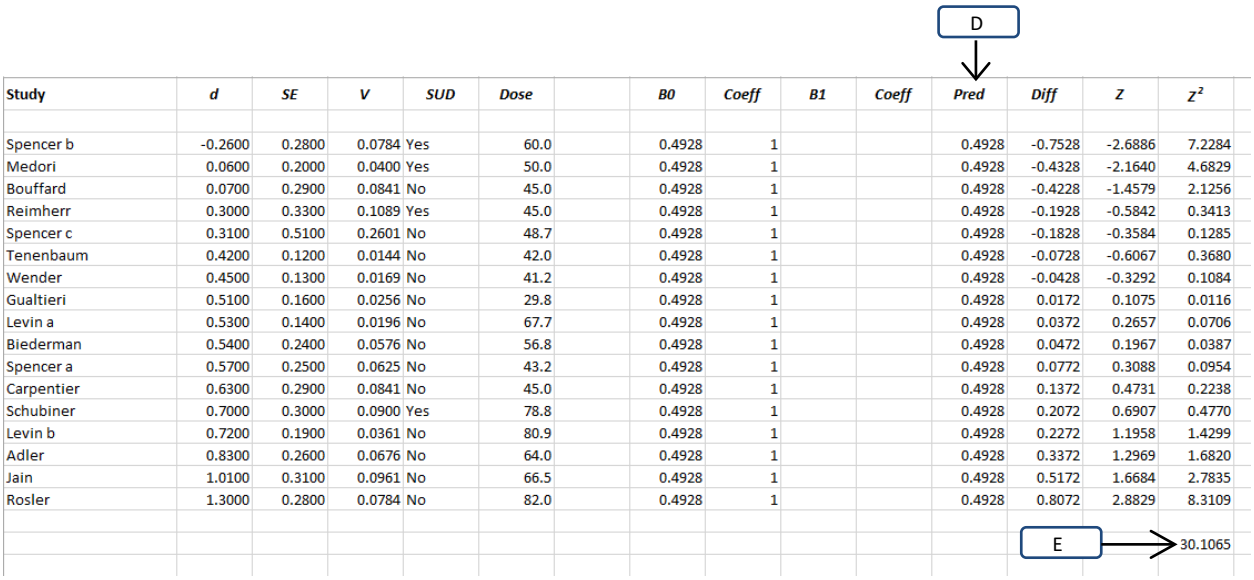


Figure 129

Using *Q* to test for heterogeneity

The null hypothesis is that all studies with the same predicted value share a common effect size, and all deviations from the predicted value are due to sampling error alone. In this example this translates to "All studies share a common effect size (the mean), and all deviations from the mean are due to sampling error alone." Under the null hypothesis *Q* would follow a chi-square distribution with degrees of freedom equal (here) to the number of studies minus one. If the test of *Q* yields a statistically significant result we conclude that the true effect size probably varies from study to study.

In Figure 129, the line for the residual (unexplained) variance [C] shows that *Q* = 30.1065 with 16 degrees of freedom, and the p-value is 0.0175. We reject the null, and conclude that (at least) some of the variance in observed effects is due to variation in true effect sizes. If we are using the fixed-effect model we conclude that the assumption that all studies share a common effect size (which is required for the fixed-effect model) has been violated.



Study	d	SE	V	SUD	Dose		B0	Coeff	B1	Coeff	Pred	Diff	Z	Z^2
Spencer b	-0.2600	0.2800	0.0784	Yes	60.0		0.4928	1			0.4928	-0.7528	-2.6886	7.2284
Medori	0.0600	0.2000	0.0400	Yes	50.0		0.4928	1			0.4928	-0.4328	-2.1640	4.6829
Bouffard	0.0700	0.2900	0.0841	No	45.0		0.4928	1			0.4928	-0.4228	-1.4579	2.1256
Reimherr	0.3000	0.3300	0.1089	Yes	45.0		0.4928	1			0.4928	-0.1928	-0.5842	0.3413
Spencer c	0.3100	0.5100	0.2601	No	48.7		0.4928	1			0.4928	-0.1828	-0.3584	0.1285
Tenenbaum	0.4200	0.1200	0.0144	No	42.0		0.4928	1			0.4928	-0.0728	-0.6067	0.3680
Wender	0.4500	0.1300	0.0169	No	41.2		0.4928	1			0.4928	-0.0428	-0.3292	0.1084
Gualtieri	0.5100	0.1600	0.0256	No	29.8		0.4928	1			0.4928	0.0172	0.1075	0.0116
Levin a	0.5300	0.1400	0.0196	No	67.7		0.4928	1			0.4928	0.0372	0.2657	0.0706
Biederman	0.5400	0.2400	0.0576	No	56.8		0.4928	1			0.4928	0.0472	0.1967	0.0387
Spencer a	0.5700	0.2500	0.0625	No	43.2		0.4928	1			0.4928	0.0772	0.3088	0.0954
Carpentier	0.6300	0.2900	0.0841	No	45.0		0.4928	1			0.4928	0.1372	0.4731	0.2238
Schubiner	0.7000	0.3000	0.0900	Yes	78.8		0.4928	1			0.4928	0.2072	0.6907	0.4770
Levin b	0.7200	0.1900	0.0361	No	80.9		0.4928	1			0.4928	0.2272	1.1958	1.4299
Adler	0.8300	0.2600	0.0676	No	64.0		0.4928	1			0.4928	0.3372	1.2969	1.6820
Jain	1.0100	0.3100	0.0961	No	66.5		0.4928	1			0.4928	0.5172	1.6684	2.7835
Rosler	1.3000	0.2800	0.0784	No	82.0		0.4928	1			0.4928	0.8072	2.8829	8.3109

Figure 130

The actual computation of *Q* is shown in in Figure 130. The predicted effect size for all studies [D] is given by the prediction equation

$$Y = 0.4928\,. \tag{1.86}$$

To compute *Q*, we work with the deviation of each study from its predicted value. For example, for the first study (Spencer b) the observed effect size is −0.2600, the predicted effect size is 0.4928, and the difference is −0.7528. We divide this by the standard error (0.2800) to get the *Z*-score (−2.6886). Then we square this to get Z^2 (7.2284). We follow the same procedure for every study, and then sum the Z^2 values to get *Q*, which is 30.10645 [E].

Random-effects

If we are using the random-effects model in the analysis, we would have used fixed-effect weights to compute *Q*, but now proceed to the random-effects tab to compute various indices of dispersion. In Figure 131 click the tab for random effects [F] and these statistics are displayed on line [G]. (Under the fixed-effect model the variation in true effects is assumed to be zero, and so those computations have no meaning.)

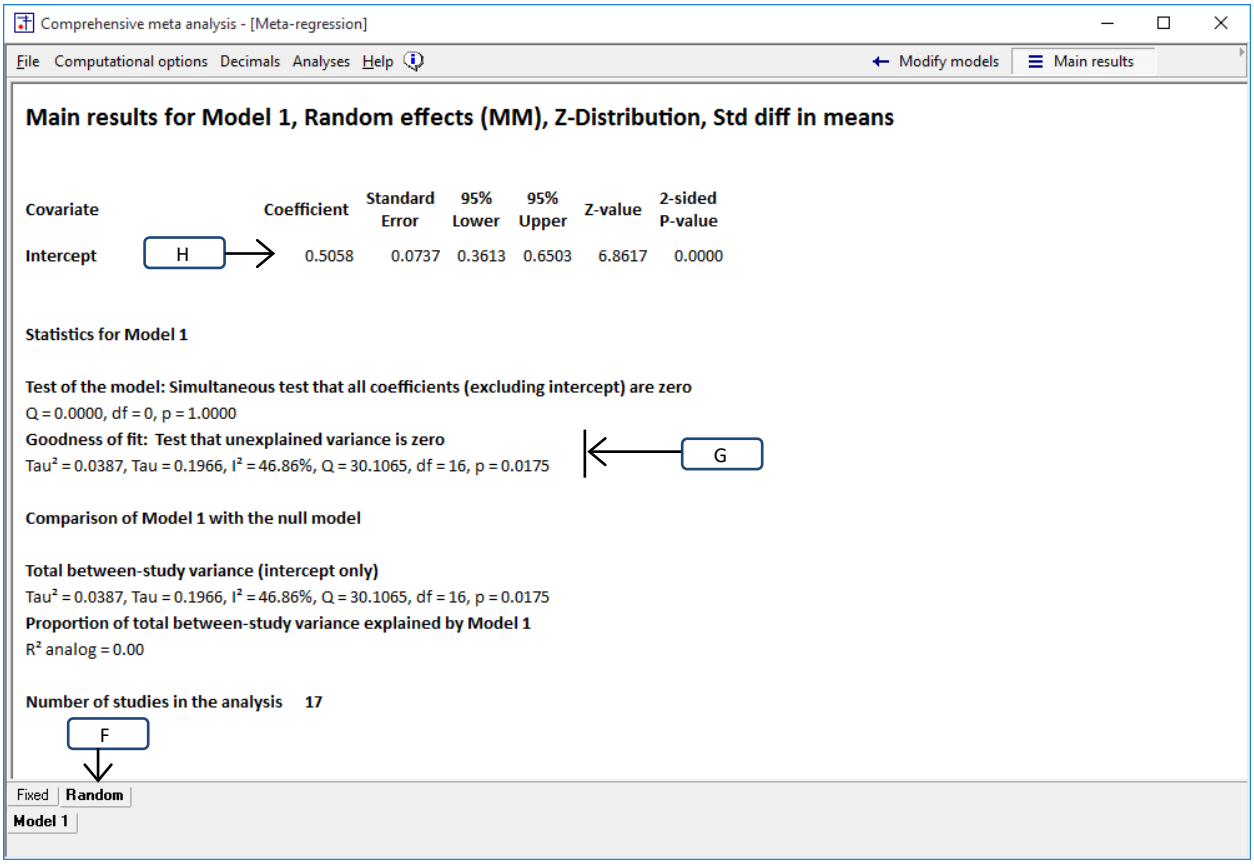


Figure 131

Using *Q* to compute the variance

In a primary study we compute the variance using

$$S^2 = \frac{SS}{df}. \tag{1.87}$$

We take a similar approach in meta-analysis, but we use the formula

$$T^2 = \frac{Q - df}{C}. \tag{1.88}$$

In the numerator, *Q* reflects the total *WSS* and *df* is the expected WSS due to sampling error. The difference, *Q* – *df*, is the part of *Q* that is attributed to variance in true effects. In the denominator, *C* is a

conversion factor that allows us to move from the standardized scale of *Q* to the metric of the effect size. In a simple analysis, *df* is the number of studies minus 1.

In this example *C* is 364.8156,

$$T^2 = \frac{30.1065 - 16}{364.8156} = 0.0387 \tag{1.89}$$

and

$$T = \sqrt{T^2} = \sqrt{0.0387} = 0.1966 \tag{1.90}$$

In Figure 131 these values are displayed on line [G].

If we assume that the true effects are normally distributed (and that our estimates of τ and of the mean effect size are accurate), then some 95% of all studies would have true effects in the range given the predicted value plus/minus 2*T*.

In Case A the predicted effect size for all studies is simply the intercept, which (using random-effects weights) is 0.5058. [H]. So, most true effects would fall in the range of

$$PRED_{LL} = 0.5058 - 2 \times 0.1966 = 0.1125 \tag{1.91}$$

$$PRED_{UL} = 0.5058 + 2 \times 0.1966 = 0.8991 \tag{1.92}$$

These are the values we use to create the graphic Figure 132. The normal curve is centered at 0.5058 and extends roughly from 0.1125 to 0.8991, intended to reflect the true effect size in some 95% of relevant studies.

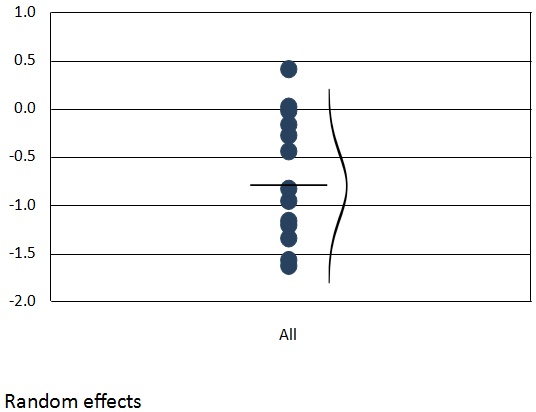


Figure 132 | Case-A | Dispersion of effects about regression line

Using *Q* to compute I^2

Recall that *Q* reflects the variance of observed effects, which incorporates (a) the variance of true effects and (b) the variance due to sampling error. Sometimes it's helpful to know what proportion of the observed variance reflects the former rather than the latter. If we call this proportion I^2, then

$$I^2 = \frac{Q_{TRUE}}{Q_{TOTAL}} \tag{1.93}$$

Since we estimate Q_{TRUE} as *Q* minus *df*, this becomes

$$I^2 = \frac{Q - df}{Q} \tag{1.94}$$

In this example

$$I^2 = \frac{30.1065 - 16}{16} = 46.86\% \tag{1.95}$$

This value is displayed on line [G]. It tells us that some 47% of the variance in observed effects about the predicted value reflects variation in true effects, while the other 53% reflects sampling error. Put another way, if each study had an extremely large sample size (and only trivial sampling error) the observed effects would fall closer to the grand mean, and the variance of these effects would be about 47% as large as the variance we are seeing now.

Case B | Subgroup analysis

Now, consider the case of a subgroups analysis. We want to use SUD to predict effect size, so we run a regression where the covariate is SUD. On the tab for fixed-effect weights [A] the prediction equation [B] is shown as 0.5504 for non-SUD studies, and as 0.5504 – 0.4003 for SUD studies. Note that we use the fixed-effect analysis to compute *Q* even if we will be using random-effects weights in the subsequent analysis.

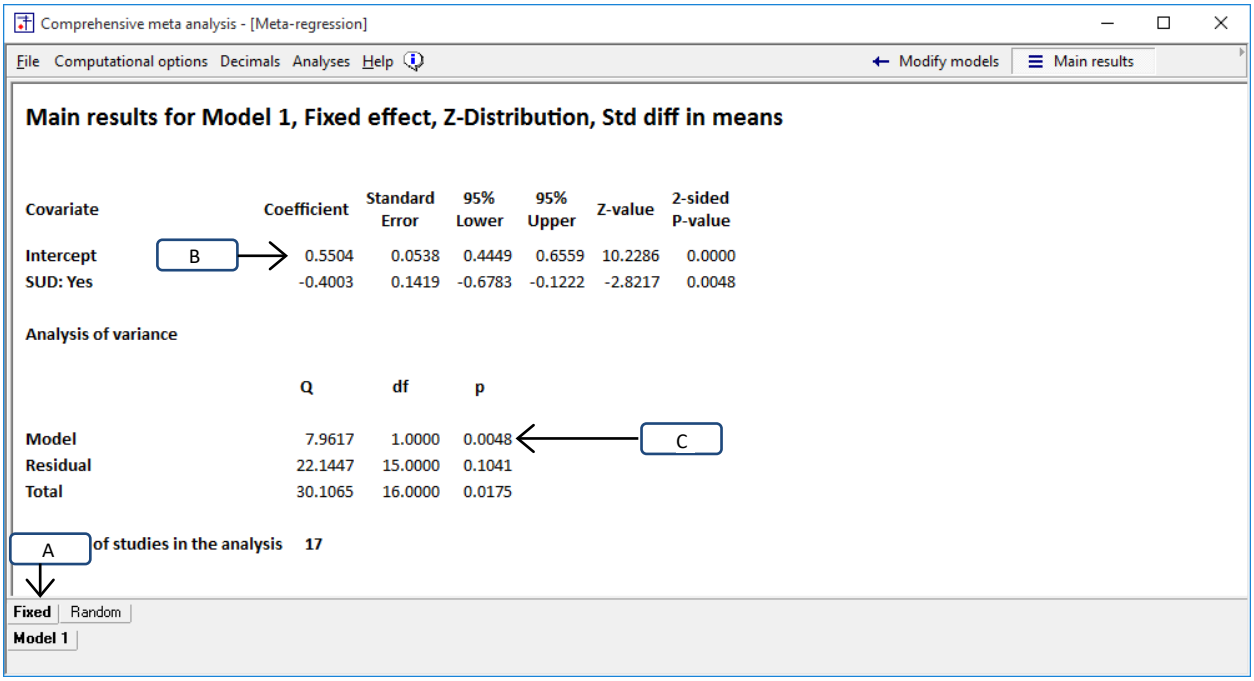
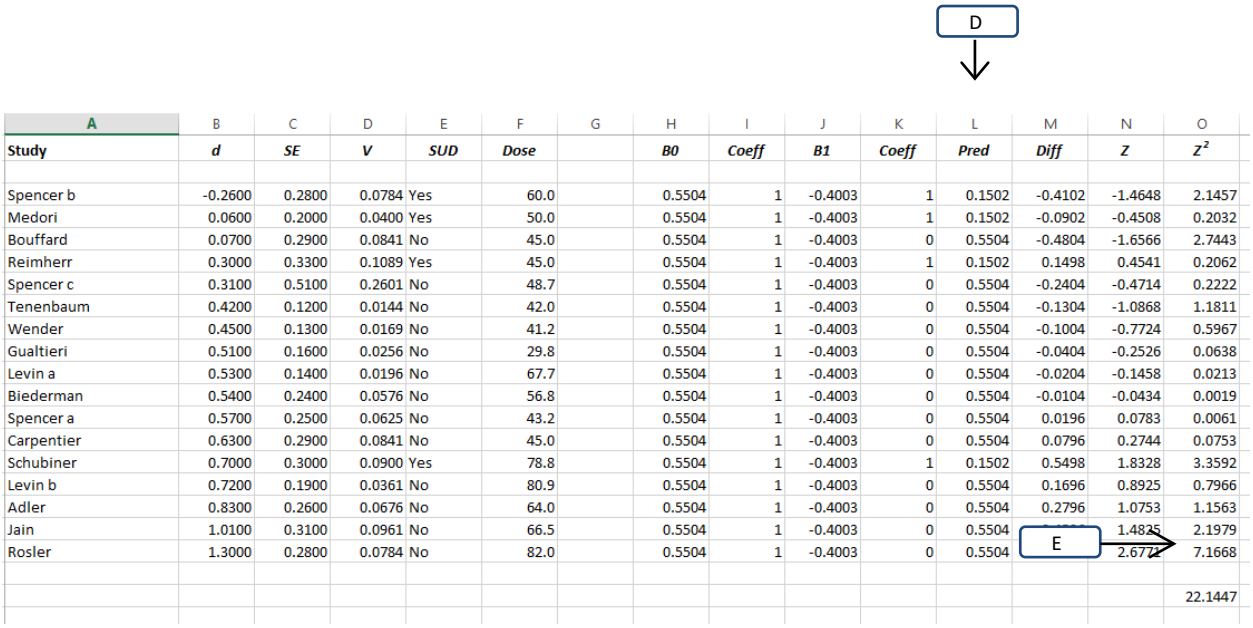


Figure 133

Using *Q* to test for heterogeneity

The null hypothesis is that all studies with the same predicted value share a common effect size, and all deviations from the predicted value are due to sampling error alone. In this example this translates to "All studies within a subgroup share a common effect size (the subgroup mean), and all deviations from this mean are due to sampling error alone." Under the null hypothesis *Q* would follow a chi-square distribution with degrees of freedom equal (here) to the number of studies minus the number of subgroups (see discussion at end of this appendix). If the test of *Q* yields a statistically significant result we conclude that the true effect size probably varies from study to study within a subgroup.

In Figure 133, the line for the residual (unexplained) variance [C] shows that *Q* = 22.1447 15 degrees of freedom, and the p-value is 0.1041. We reject the null, and conclude that (at least) some of the variance in observed effects about the subgroup means is due to variation in true effect sizes. If we are using the fixed-effect model we conclude that the assumption that all studies within a subgroup share a common effect size (which is required for the fixed-effect model) has been violated.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Study	***d***	***SE***	***V***	***SUD***	***Dose***		***B0***	***Coeff***	***B1***	***Coeff***	***Pred***	***Diff***	***Z***	Z^2
Spencer b	-0.2600	0.2800	0.0784	Yes	60.0		0.5504	1	-0.4003	1	0.1502	-0.4102	-1.4648	2.1457
Medori	0.0600	0.2000	0.0400	Yes	50.0		0.5504	1	-0.4003	1	0.1502	-0.0902	-0.4508	0.2032
Bouffard	0.0700	0.2900	0.0841	No	45.0		0.5504	1	-0.4003	0	0.5504	-0.4804	-1.6566	2.7443
Reimherr	0.3000	0.3300	0.1089	Yes	45.0		0.5504	1	-0.4003	1	0.1502	0.1498	0.4541	0.2062
Spencer c	0.3100	0.5100	0.2601	No	48.7		0.5504	1	-0.4003	0	0.5504	-0.2404	-0.4714	0.2222
Tenenbaum	0.4200	0.1200	0.0144	No	42.0		0.5504	1	-0.4003	0	0.5504	-0.1304	-1.0868	1.1811
Wender	0.4500	0.1300	0.0169	No	41.2		0.5504	1	-0.4003	0	0.5504	-0.1004	-0.7724	0.5967
Gualtieri	0.5100	0.1600	0.0256	No	29.8		0.5504	1	-0.4003	0	0.5504	-0.0404	-0.2526	0.0638
Levin a	0.5300	0.1400	0.0196	No	67.7		0.5504	1	-0.4003	0	0.5504	-0.0204	-0.1458	0.0213
Biederman	0.5400	0.2400	0.0576	No	56.8		0.5504	1	-0.4003	0	0.5504	-0.0104	-0.0434	0.0019
Spencer a	0.5700	0.2500	0.0625	No	43.2		0.5504	1	-0.4003	0	0.5504	0.0196	0.0783	0.0061
Carpentier	0.6300	0.2900	0.0841	No	45.0		0.5504	1	-0.4003	0	0.5504	0.0796	0.2744	0.0753
Schubiner	0.7000	0.3000	0.0900	Yes	78.8		0.5504	1	-0.4003	1	0.1502	0.5498	1.8328	3.3592
Levin b	0.7200	0.1900	0.0361	No	80.9		0.5504	1	-0.4003	0	0.5504	0.1696	0.8925	0.7966
Adler	0.8300	0.2600	0.0676	No	64.0		0.5504	1	-0.4003	0	0.5504	0.2796	1.0753	1.1563
Jain	1.0100	0.3100	0.0961	No	66.5		0.5504	1	-0.4003	0	0.5504		1.4825	2.1979
Rosler	1.3000	0.2800	0.0784	No	82.0		0.5504	1	-0.4003	0	0.5504		2.6772	7.1668
														22.1447

Figure 134

The actual computation of *Q* is shown in in Figure 134. The predicted effect size for each study is

$$Y = 0.5504 - 0.4003 \times SUD\,. \tag{1.96}$$

To compute *Q*, we work with the deviation of each study from its predicted value. For example, for the first study (Spencer b) the observed effect size is −0.2600, the predicted effect size is 0.15015, and the difference is −0.4102. We divide this by the standard error (0.2800) to get the *Z*-score (−1.46484). Then we square this to get Z^2 (2.145742). We follow the same procedure for every study, and then sum the Z^2 values to get *Q*, which is 22.14475.

Random-effects

If we are using the random-effects model in the analysis, we would have used fixed-effect weights to compute *Q*, but now proceed to the random-effects tab to compute various indices of dispersion. In Figure 131 click the tab for random effects [F] and these statistics are displayed on line [G]. (Under the fixed-effect model the variation in true effects is assumed to be zero, and so those computations have no meaning.)

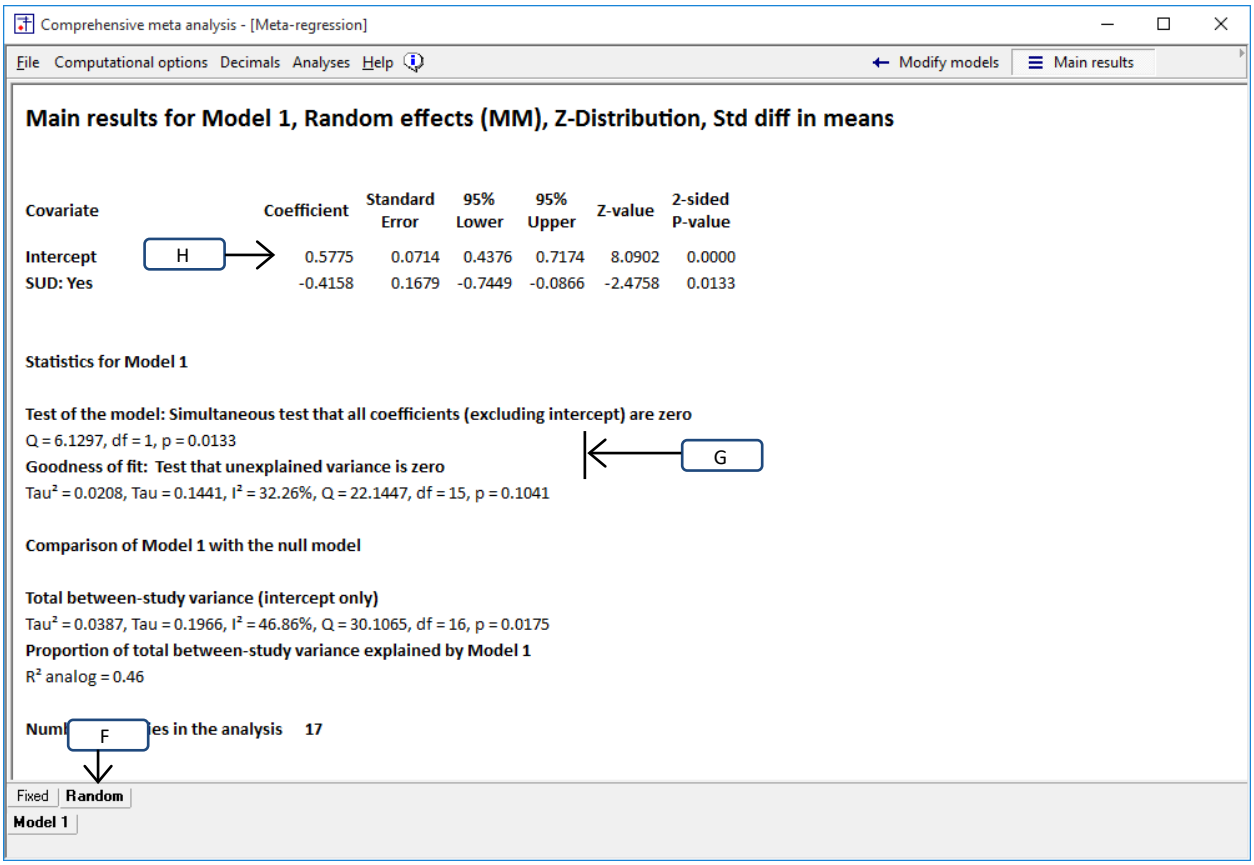


Figure 135

Using *Q* to compute the variance

As before, we compute T2 using

$$T^2 = \frac{Q - df}{C}. \tag{1.97}$$

Here, *C* is 344.0653 and so

$$T^2 = \frac{22.1447 - 15}{344.0653} = 0.0208 \tag{1.98}$$

and

$$T = \sqrt{T^2} = \sqrt{0.0208} = 0.1441 \tag{1.99}$$

If we assume that the true effects are normally distributed about their predicted value then some 95% of all studies would have true effects in the range given the predicted value plus/minus 2*T*.

Under the random-effects weights the predicted effect for the non-SUD studies 0.5775, and so the true effects would fall in the range of

$$LL = 0.5775 - 2 \times 0.1442 = 0.2891 \tag{1.100}$$

$$LL = 0.5775 + 2 \times 0.1442 = 0.8659\,. \tag{1.101}$$

Similarly, the predicted effect for the SUD studies 0.1617, and so the true effects would fall in the range of

$$LL = 0.1617 - 2 \times 0.1442 = -0.1267 \tag{1.102}$$

$$LL = 0.1617 + 2 \times 0.1442 = 0.4501\,. \tag{1.103}$$

These are the values we use to create the graphic. The normal curve for non-SUD is centered at 0.5775 and extends roughly from 0.2891 to 0.8659, intended to reflect the true effect size in some 95% of relevant studies. The normal curve for SUD is centered at 0.1617 and extends roughly from −0.1267 to 0.4501, intended to reflect the true effect size in some 95% of relevant studies.

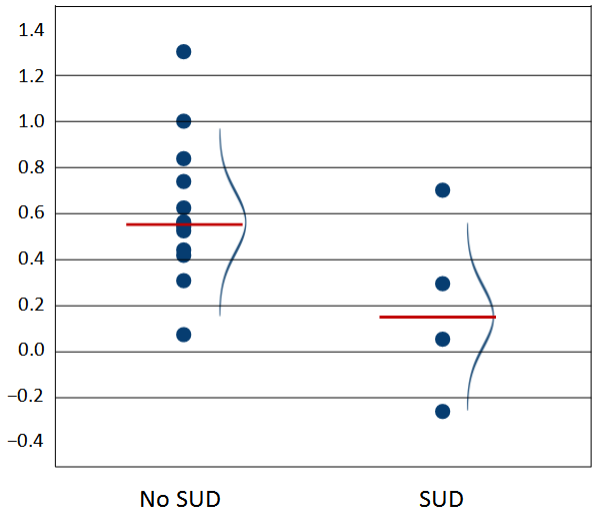


Figure 136 | Dispersion of effects about the subgroup means

Using Q to compute I^2

As before, we compute

$$I^2 = \frac{Q - df}{Q} \tag{1.104}$$

In this example

$$I^2 = \left(\frac{22.1447 - 15}{15}\right) \times 100 = 32.2639\% \tag{1.105}$$

This value is displayed on line [G]. It tells us that some 32% of the variance in observed effects about either subgroup mean reflects variation in true effects, while the other 68% reflects sampling error. Put another way, if each study had an extremely large sample size (and only trivial sampling error) the observed points in each subgroup would fall closer to their subgroup mean. The variance of these effects would be about 32% as large as the variance we are seeing now.

Case C | Continuous covariate

Finally, consider the case where we want to estimate the effect size as a function of a continuous covariate such as Dose. We run the regression with Dose as a covariate. On the tab for fixed-effect weights [A] the prediction equation [B] is shown as 0.0814 + 0.0079 x Dose. Note that we use the fixed-effect analysis to compute *Q* even if we will be using random-effects weights in the subsequent analysis.

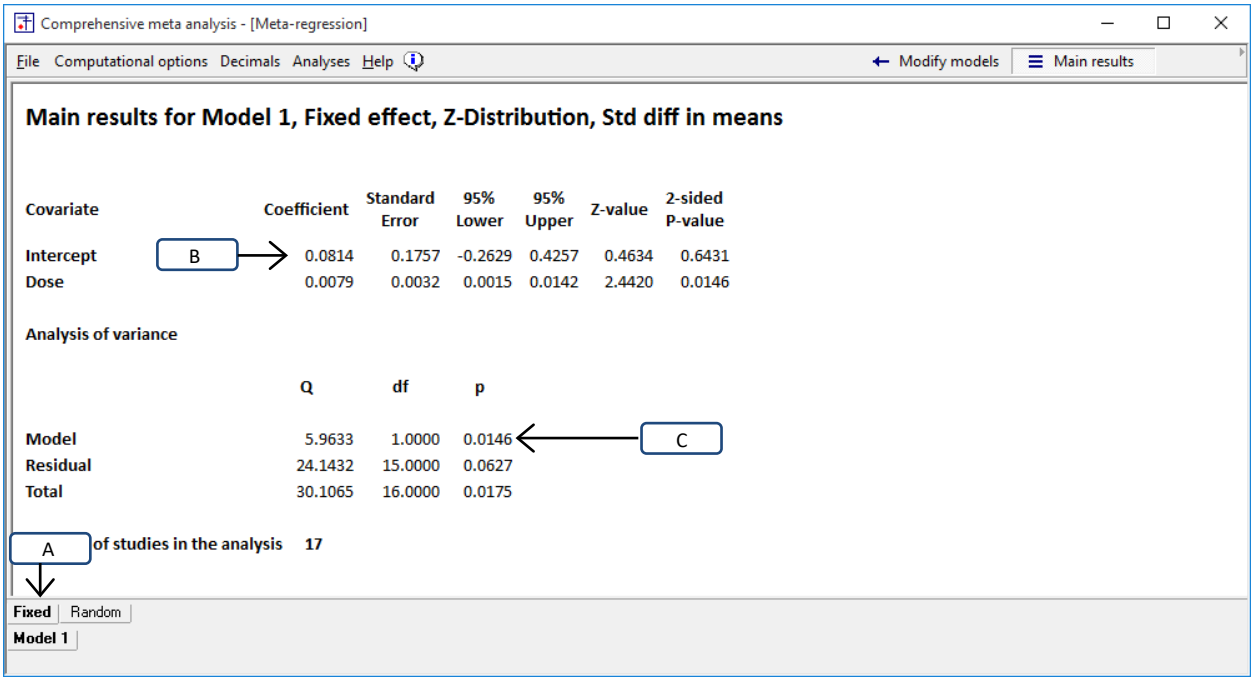
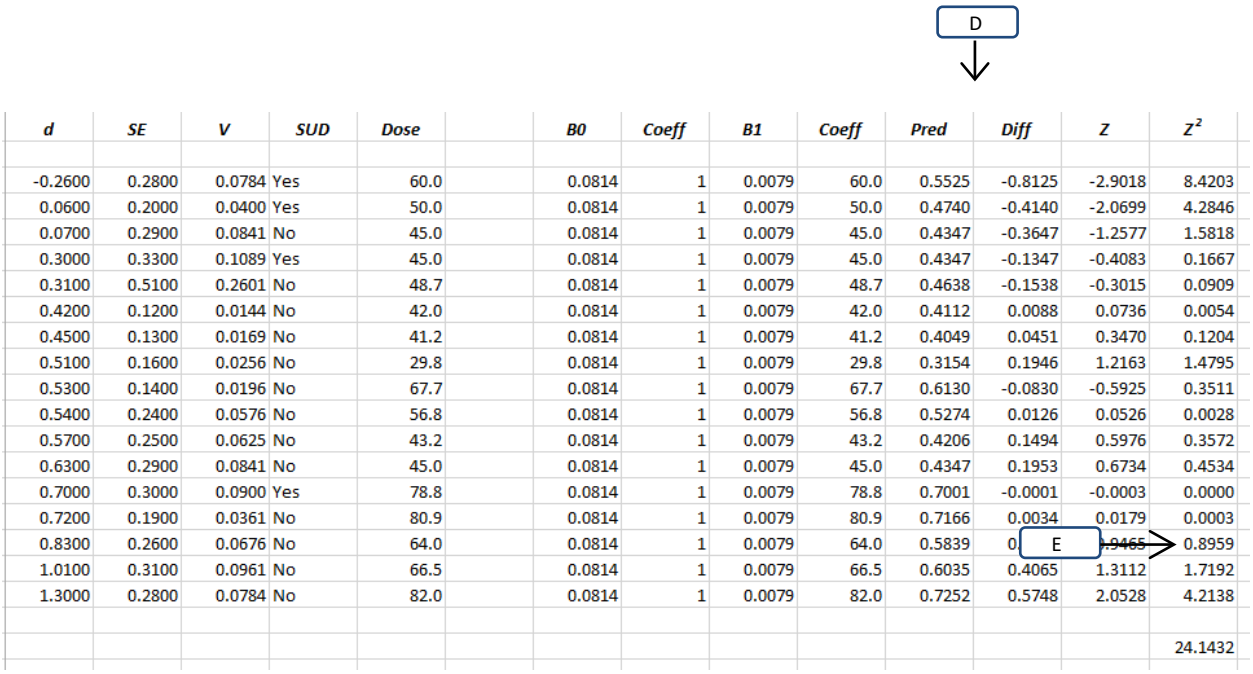


Figure 137

Using *Q* to test for heterogeneity

Under the null hypothesis that all studies with the same predicted value share a common effect size, and that all deviations from the regression line are due to sampling error alone, *Q* would follow a chi-square distribution with degrees of freedom equal to the number of studies minus 1, minus the number of covariates. So, we can use *Q* to test the null hypothesis. If the test of *Q* yields a statistically significant result we can say that the null is probably false, and the true effect size probably varies from study to study for any given point on the regression line.

In Figure 137, the line for the residual (unexplained) variance [C] shows that *Q* = 24.1432 with 15 degrees of freedom, and the p-value is 0.0627. We reject the null, and conclude that some of the variance in observed effects about the regression line is due to variation in true effect sizes. If we are using the fixed-effect model we conclude that the assumption that all studies with the same predicted value share a common effect size (which is required for the fixed-effect model) has been violated.



d	SE	V	SUD	Dose		B0	Coeff	B1	Coeff	Pred	Diff	Z	Z^2
-0.2600	0.2800	0.0784	Yes	60.0		0.0814	1	0.0079	60.0	0.5525	-0.8125	-2.9018	8.4203
0.0600	0.2000	0.0400	Yes	50.0		0.0814	1	0.0079	50.0	0.4740	-0.4140	-2.0699	4.2846
0.0700	0.2900	0.0841	No	45.0		0.0814	1	0.0079	45.0	0.4347	-0.3647	-1.2577	1.5818
0.3000	0.3300	0.1089	Yes	45.0		0.0814	1	0.0079	45.0	0.4347	-0.1347	-0.4083	0.1667
0.3100	0.5100	0.2601	No	48.7		0.0814	1	0.0079	48.7	0.4638	-0.1538	-0.3015	0.0909
0.4200	0.1200	0.0144	No	42.0		0.0814	1	0.0079	42.0	0.4112	0.0088	0.0736	0.0054
0.4500	0.1300	0.0169	No	41.2		0.0814	1	0.0079	41.2	0.4049	0.0451	0.3470	0.1204
0.5100	0.1600	0.0256	No	29.8		0.0814	1	0.0079	29.8	0.3154	0.1946	1.2163	1.4795
0.5300	0.1400	0.0196	No	67.7		0.0814	1	0.0079	67.7	0.6130	-0.0830	-0.5925	0.3511
0.5400	0.2400	0.0576	No	56.8		0.0814	1	0.0079	56.8	0.5274	0.0126	0.0526	0.0028
0.5700	0.2500	0.0625	No	43.2		0.0814	1	0.0079	43.2	0.4206	0.1494	0.5976	0.3572
0.6300	0.2900	0.0841	No	45.0		0.0814	1	0.0079	45.0	0.4347	0.1953	0.6734	0.4534
0.7000	0.3000	0.0900	Yes	78.8		0.0814	1	0.0079	78.8	0.7001	-0.0001	-0.0003	0.0000
0.7200	0.1900	0.0361	No	80.9		0.0814	1	0.0079	80.9	0.7166	0.0034	0.0179	0.0003
0.8300	0.2600	0.0676	No	64.0		0.0814	1	0.0079	64.0	0.5839	0	-.9465	0.8959
1.0100	0.3100	0.0961	No	66.5		0.0814	1	0.0079	66.5	0.6035	0.4065	1.3112	1.7192
1.3000	0.2800	0.0784	No	82.0		0.0814	1	0.0079	82.0	0.7252	0.5748	2.0528	4.2138
													24.1432

Figure 138

The actual computation of *Q* is shown in Figure 138. The predicted effect size for each study is

$$Y = 0.0814 + 0.0079 \times Dose\,. \tag{1.106}$$

To compute *Q*, we work with the deviation of each study from its predicted value. For the first study (Spencer b) the observed effect size is −0.2600, the predicted effect size is 0.5525, and the difference is −0.8125. We divide this by the standard error (0.2800) to get the *Z*-score (−2.9018). Then we square this to get Z^2 (8.4203). We follow the same procedure for every study, and then sum the Z^2 values to get *Q*, which is 24.1432.

Random-effects

If we are using the random-effects model in the analysis, we would have used fixed-effect weights to compute *Q*, but now proceed to the random-effects tab to compute various indices of dispersion. In Figure 131 click the tab for random effects [F] and these statistics are displayed on line [G]. (Under the fixed-effect model the variation in true effects is assumed to be zero, and so those computations have no meaning.)

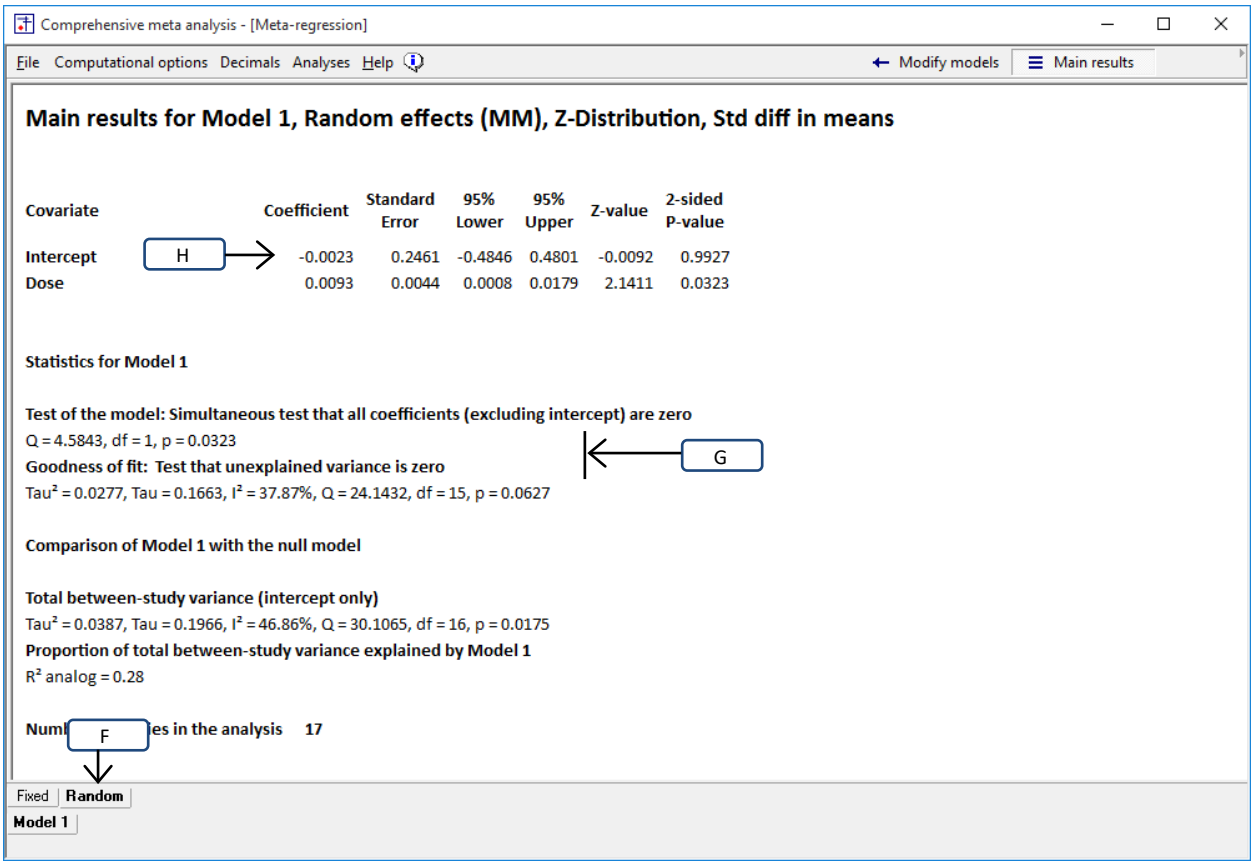


Figure 139

Using *Q* to compute the variance

As before, we compute T2 using

$$T^2 = \frac{Q - df}{C}. \tag{1.107}$$

In this example C is _____,

$$T^2 = \frac{24.1432 - 15}{??} = 0.0277 \tag{1.108}$$

and

$$T = \sqrt{T^2} = \sqrt{0.0277} = 0.1664 \tag{1.109}$$

If we assume that the true effects are normally distributed about their predicted value (and that the predicted value is correct), then 95% of all studies would have true effects in the range given the predicted value plus/minus approximately 2*T*.

For example, for studies with a dose of 30 the predicted effect size is −0.0023 +0.0093 x 30 = 0.3170. If this estimate is correct, then most studies with a dose of 30 would have a true effect size in the range of

$$LL = 0.3170 - 2 \times 0.1664 = -0.0159 \tag{1.110}$$

$$LL = 0.3170 + 2 \times 0.1664 = 0.6499 \tag{1.111}$$

And, studies with a dose of 80 the predicted effect size is −0.0023 +0.0093 x 80 = 0.7095. If this estimate is correct, then most studies with a dose of 80 would have a true effect size in the range of

$$LL = 0.7095 - 2 \times 0.1664 = 0.3766 \tag{1.112}$$

$$LL = 0.7095 + 2 \times 0.1664 = 1.0424 \tag{1.113}$$

These are the values we use to create the graphic. For a dose of 30 the normal curve is centered at 0.3170 and extends roughly from −0.0159 to 0.6499, intended to reflect the true effect size in some 95% of relevant studies. For a dose of 80 the normal curve is centered at 0.7095 and extends roughly from −0.3766 to 1.0424, intended to reflect the true effect size in some 95% of relevant studies. (This simplified example assumes that the mean is known. To actually compute prediction intervals we would take account of the error in estimating the mean.)

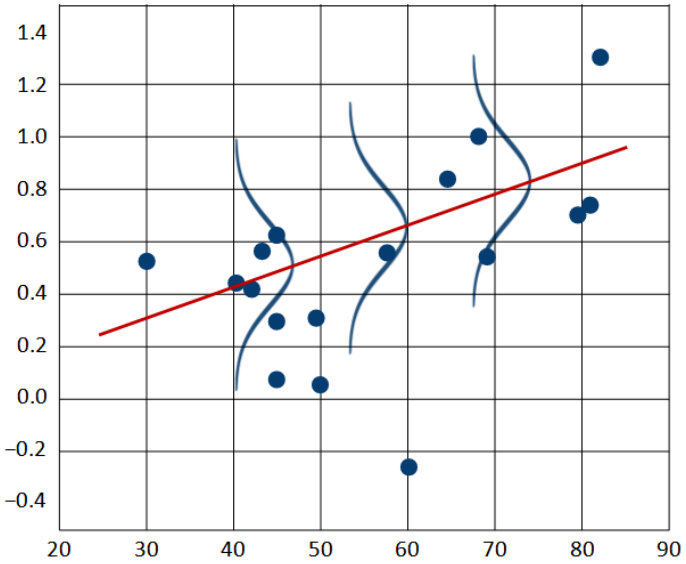


Figure 140 | Dispersion of effects about regression line for dose

Using Q to compute I^2

As before, we compute

$$I^2 = \frac{Q - df}{Q} \tag{1.114}$$

In this example

$$I^2 = \frac{24.1432 - 15}{15} = 37.87\% \tag{1.115}$$

This value is displayed on line [G]. It tells us that some 38% of the variance in observed effects at any point on the regression line reflects variation in true effects, while the other 62% reflects sampling error. Put another way, if each study had an extremely large sample size (and only trivial sampling error) the observed points would fall closer to the regression line. The variance of these effects would be about 40% as large as the variance we are seeing now.

To this point we've presented an example for a simple analysis, for a subgroups analysis, and for an analysis where we used a continuous predictor. We presented each analysis using the same format to highlight the fact that the same formulas apply in all three cases.

That is, we compute *Q* by working the deviations of each observed effect size from the predicted effect size. In Case A the predicted effect size is simply the intercept, while in Case B or C it involves two or more coefficients. However, all three cases are functionally the same, and this can be extended to analyses with any number of covariates.

The degrees of freedom in Case A was given as the number of studies minus 1, in Case B as the number of studies minus the number of subgroups, and in Case C as the number of studies minus 1 minus the number of covariates. The general rule is that the degrees of freedom is equal to the number of studies minus the number of predictors, where the intercept or a covariate count as predictors.

We showed how to compute the approximate range of effects using the predicted value plus or minus 2T. This assumes that the predicted effect is accurate, and that our estimate of T is accurate. If we actually wanted to compute a prediction interval we would probably want to adjust the interval to take account of the fact that both the predicted value and T are estimated with error.

APPENDIX 4: COMPUTING τ^2 IN THE PRESENCE OF SUBGROUPS

When we're working with subgroups, it's clear that we need to estimate the between-study variance T2 within subgroups rather than for the full set of studies. In our example, when we're estimating the mean effect for Hot studies and for Cold studies, the between-study variance that we need to assign weights and to discuss the unexplained variance is clearly the variance within subgroups.

However, there are two ways to estimate τ^2 within subgroups.

One option is to compute one estimate of τ^2 for the Hot studies, and a separate estimate for the Cold studies. Then, we would use each estimate for the corresponding set of studies.

The other option is to compute one estimate of τ^2 for the Hot studies and a separate estimate for the Cold studies. Then we would pool the two estimates and use the pooled estimate for both sets of studies.

The logic for choosing the second option is that estimates of τ^2 are not reliable unless they're based on a large number of studies. Unless we have good reason to believe that τ^2 will differ substantially from one subgroup to the next, it's often a better idea to assume that the true value of τ^2 is comparable for each subgroup, and we'll get a better estimate of this common value by pooling the within-subgroups estimates.

This is the option that we use with meta-regression since (at least when we're using continuous covariates) the first option is not tenable. For purposes of comparing the subgroups analyses with the regression, we must select this option for subgroups as well.

On the analysis screen

- Select Computational options > Mixed and random effects options
- Select the option to Assume a common among-study variance across subgroups
- Select the option to Combine subgroup using a fixed-effect model

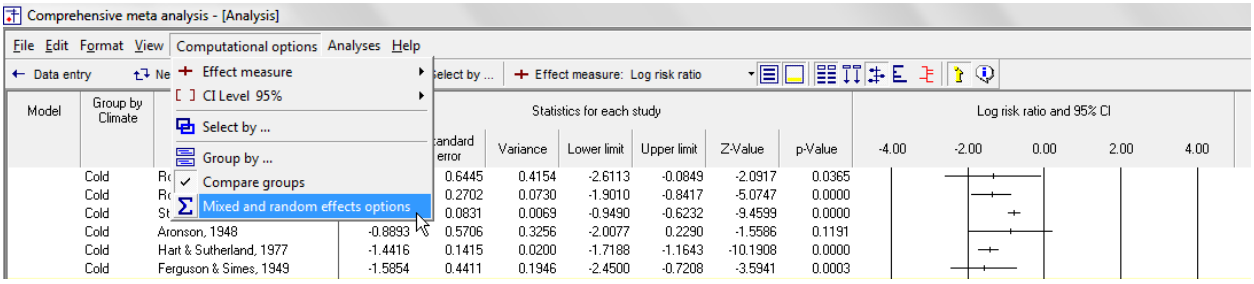


Figure 141 | Computing τ^2 in the presence of subgroups

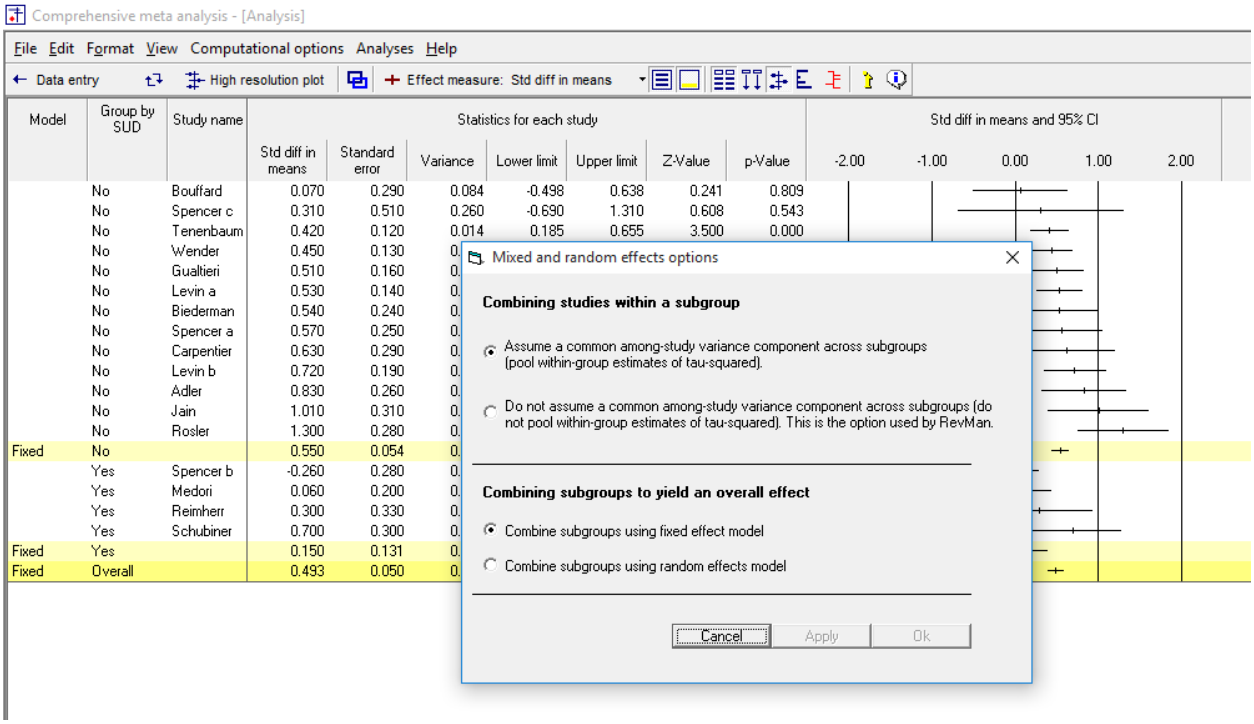


Figure 142 | Computing τ^2 in the presence of subgroups

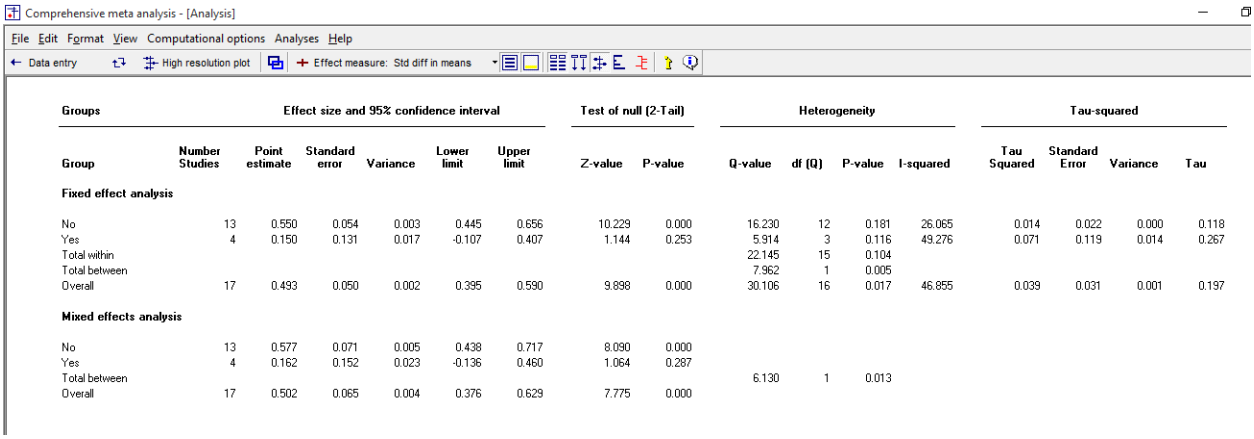


Figure 143 | Computing τ^2 in the presence of subgroups

In this example we would pool the within-subgroup estimates of T2 and apply the pooled estimate to both subgroups. The program does not display the pooled estimate on this screen. To see the pooled estimate, click Next table and Calculations. The pooled estimate is 0.0964.

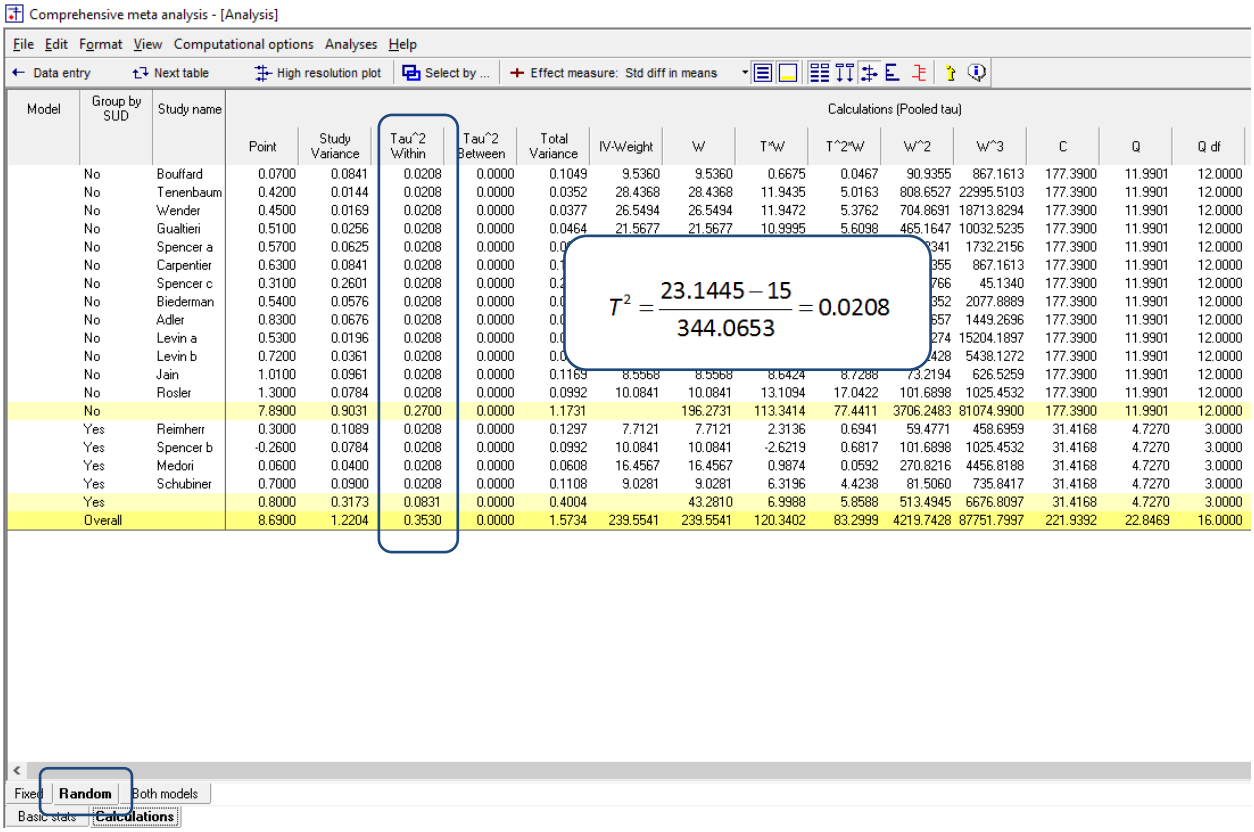


Figure 144 | Computing τ^2 in the presence of subgroups

For those who are interested, we show how to actually pool the estimates of T^2 to get the pooled estimate, and how to display the pooled estimate. The mechanism for pooling is not to take the mean of the two estimates, but rather to pool the underlying statistics and then computed T^2 from the combined data.

Recall that

$$T^2 = \frac{Q - df}{C} \tag{1.116}$$

and so to compute a pooled value we use

$$T^2 = \frac{\sum Q - \sum df}{\sum C} \tag{1.117}$$

where values are summed across all subgroups. To get the within-subgroup values prior to pooling

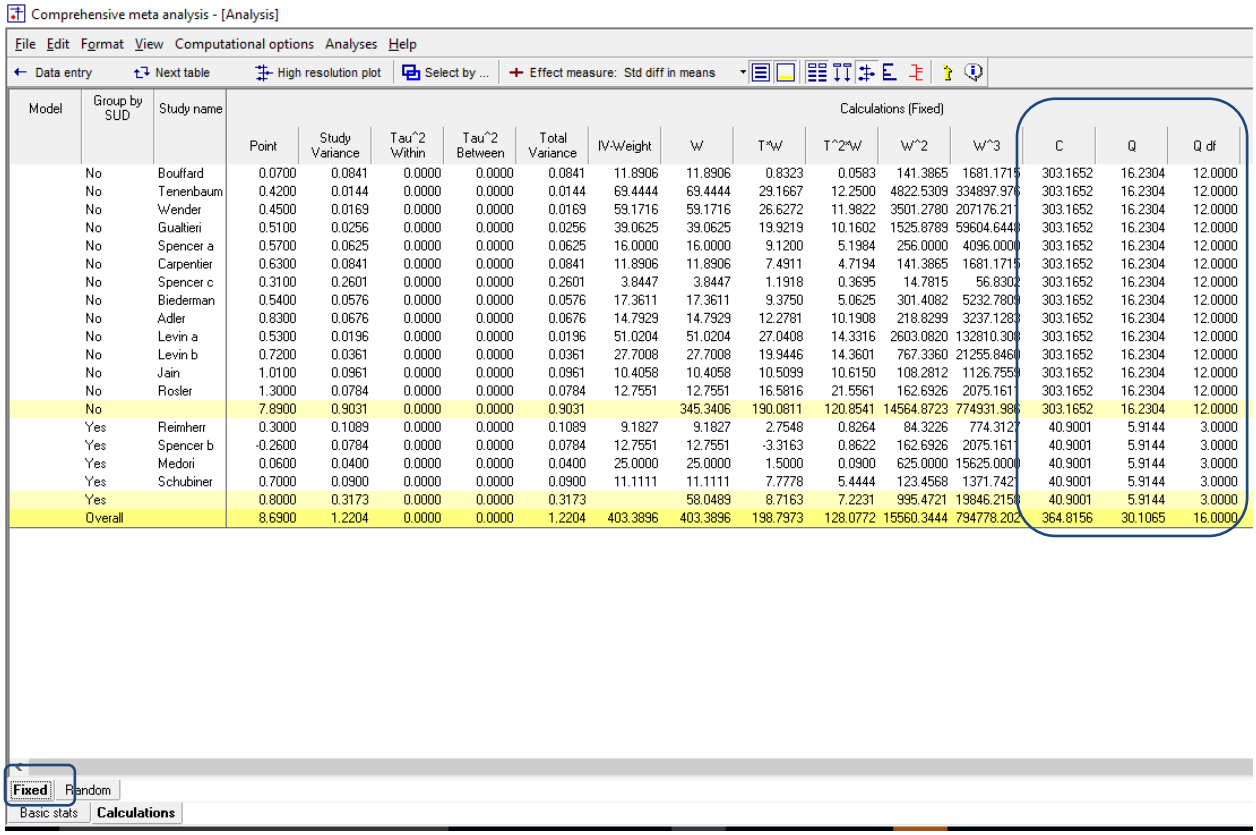
- Select “Do not assume a common among-study variance component”.
- Click Fixed

- Click Calculations

In Figure 145

- Column C shows within-subgroup values of the constant *C*
- Column *Q* shows within-subgroup values of *Q*
- Column *Q* df shows within-subgroup values of *df*

	C	*Q*	*df*
No	303.1652	16.2304	12
Yes	40.9001	5.9144	3
Sum	344.0653	22.1448	15



Model	Group by SUD	Study name	Point	Study Variance	Tau^2 Within	Tau^2 Between	Total Variance	IV-Weight	W	T*W	T^2*W	W^2	W^3	C	Q	Q df
	No	Bouffard	0.0700	0.0841	0.0000	0.0000	0.0841	11.8906	11.8906	0.8323	0.0583	141.3865	1681.1715	303.1652	16.2304	12.0000
	No	Tenenbaum	0.4200	0.0144	0.0000	0.0000	0.0144	69.4444	69.4444	29.1667	12.2500	4822.5309	334897.976	303.1652	16.2304	12.0000
	No	Wender	0.4500	0.0169	0.0000	0.0000	0.0169	59.1716	59.1716	26.6272	11.9822	3501.2780	207176.21	303.1652	16.2304	12.0000
	No	Gualtieri	0.5100	0.0256	0.0000	0.0000	0.0256	39.0625	39.0625	19.9219	10.1602	1525.8789	59604.6448	303.1652	16.2304	12.0000
	No	Spencer a	0.5700	0.0625	0.0000	0.0000	0.0625	16.0000	16.0000	9.1200	5.1984	256.0000	4096.0000	303.1652	16.2304	12.0000
	No	Carpentier	0.6300	0.0841	0.0000	0.0000	0.0841	11.8906	11.8906	7.4911	4.7194	141.3865	1681.1715	303.1652	16.2304	12.0000
	No	Spencer c	0.3100	0.2601	0.0000	0.0000	0.2601	3.8447	3.8447	1.1918	0.3695	14.7815	56.8302	303.1652	16.2304	12.0000
	No	Biederman	0.5400	0.0576	0.0000	0.0000	0.0576	17.3611	17.3611	9.3750	5.0625	301.4082	5232.7809	303.1652	16.2304	12.0000
	No	Adler	0.8300	0.0676	0.0000	0.0000	0.0676	14.7929	14.7929	12.2781	10.1908	218.8299	3237.1283	303.1652	16.2304	12.0000
	No	Levin a	0.5300	0.0196	0.0000	0.0000	0.0196	51.0204	51.0204	27.0408	14.3316	2603.0820	132810.308	303.1652	16.2304	12.0000
	No	Levin b	0.7200	0.0361	0.0000	0.0000	0.0361	27.7008	27.7008	19.9446	14.3601	767.3360	21255.8460	303.1652	16.2304	12.0000
	No	Jain	1.0100	0.0961	0.0000	0.0000	0.0961	10.4058	10.4058	10.5099	10.6150	108.2812	1126.7559	303.1652	16.2304	12.0000
	No	Rosler	1.3000	0.0784	0.0000	0.0000	0.0784	12.7551	12.7551	16.5816	21.5561	162.6926	2075.161	303.1652	16.2304	12.0000
	No		7.8900	0.9031	0.0000	0.0000	0.9031		345.3406	190.0811	120.8541	14564.8723	774931.986	303.1652	16.2304	12.0000
	Yes	Reimherr	0.3000	0.1089	0.0000	0.0000	0.1089	9.1827	9.1827	2.7548	0.8264	84.3226	774.3127	40.9001	5.9144	3.0000
	Yes	Spencer b	-0.2600	0.0784	0.0000	0.0000	0.0784	12.7551	12.7551	-3.3163	0.8622	162.6926	2075.161	40.9001	5.9144	3.0000
	Yes	Medori	0.0600	0.0400	0.0000	0.0000	0.0400	25.0000	25.0000	1.5000	0.0900	625.0000	15625.0000	40.9001	5.9144	3.0000
	Yes	Schubiner	0.7000	0.0900	0.0000	0.0000	0.0900	11.1111	11.1111	7.7778	5.4444	123.4568	1371.742	40.9001	5.9144	3.0000
	Yes		0.8000	0.3173	0.0000	0.0000	0.3173		58.0489	8.7163	7.2231	995.4721	19846.2158	40.9001	5.9144	3.0000
	Overall		8.6900	1.2204	0.0000	0.0000	1.2204	403.3896	403.3896	198.7973	128.0772	15560.3444	794778.202	364.8156	30.1065	16.0000

Figure 145 | Computing τ² in the presence of subgroups

$$T^2 = \frac{23.1445 - 15}{344.0653} = 0.0208 \qquad (1.118)$$

which is the same value we saw in Figure 144.

Be sure to switch the option back to “Assume a common value” and to re-select the tab for Random.

APPENDIX 5: CREATING VARIABLES FOR INTERACTIONS

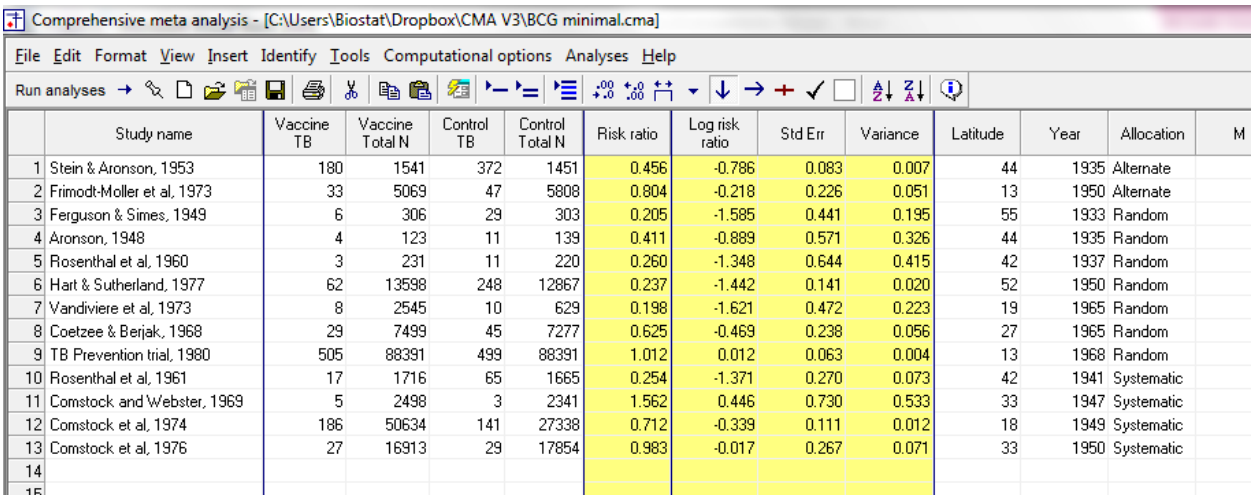
See Data files and downloads for location of files

In general, it's a good idea to create all the variables you'll need on the data-entry screen before proceeding to the analysis. However, if you're in the middle of the analyses and discover that you need to create additional variables, you can easily return to the data-entry screen to do so.

In some cases it's easiest to enter data for the new variables manually. This may be the case if you have only a few studies and the computation of the new data points is simple (for example, each study is coded 0 or 1).

In other cases, however, it's easier to copy the data out to Excel, create the data for the new variables, and then copy the data back into CMA. We show that process here.

Suppose we're working with the data set shown in Figure 146, which includes the variable Dose. At some point we realize that we need a variable called Dose-C (centered) and one called Dose-C2 (centered, then squared) for the analyses.



	Study name	Vaccine TB	Vaccine Total N	Control TB	Control Total N	Risk ratio	Log risk ratio	Std Err	Variance	Latitude	Year	Allocation	M
1	Stein & Aronson, 1953	180	1541	372	1451	0.456	-0.786	0.083	0.007	44	1935	Alternate	
2	Frimodt-Moller et al, 1973	33	5069	47	5808	0.804	-0.218	0.226	0.051	13	1950	Alternate	
3	Ferguson & Simes, 1949	6	306	29	303	0.205	-1.585	0.441	0.195	55	1933	Random	
4	Aronson, 1948	4	123	11	139	0.411	-0.889	0.571	0.326	44	1935	Random	
5	Rosenthal et al, 1960	3	231	11	220	0.260	-1.348	0.644	0.415	42	1937	Random	
6	Hart & Sutherland, 1977	62	13598	248	12867	0.237	-1.442	0.141	0.020	52	1950	Random	
7	Vandiviere et al, 1973	8	2545	10	629	0.198	-1.621	0.472	0.223	19	1965	Random	
8	Coetzee & Berjak, 1968	29	7499	45	7277	0.625	-0.469	0.238	0.056	27	1965	Random	
9	TB Prevention trial, 1980	505	88391	499	88391	1.012	0.012	0.063	0.004	13	1968	Random	
10	Rosenthal et al, 1961	17	1716	65	1665	0.254	-1.371	0.270	0.073	42	1941	Systematic	
11	Comstock and Webster, 1969	5	2498	3	2341	1.562	0.446	0.730	0.533	33	1947	Systematic	
12	Comstock et al, 1974	186	50634	141	27338	0.712	-0.339	0.111	0.012	18	1949	Systematic	
13	Comstock et al, 1976	27	16913	29	17854	0.983	-0.017	0.267	0.071	33	1950	Systematic	
14													

Figure 146 | Creating variables for interactions

- Return to this screen.

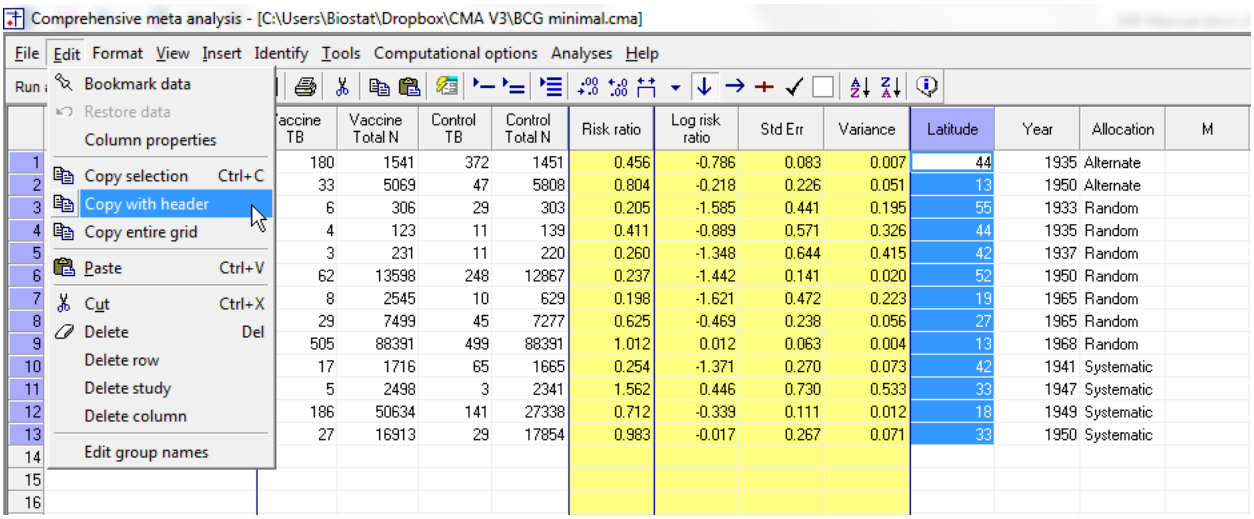


Figure 147 | Creating variables for interactions

Insert a column called Dose-C and define it as Moderator > Decimal
Insert a column called Dose-C2 and define it as Moderator > Decimal

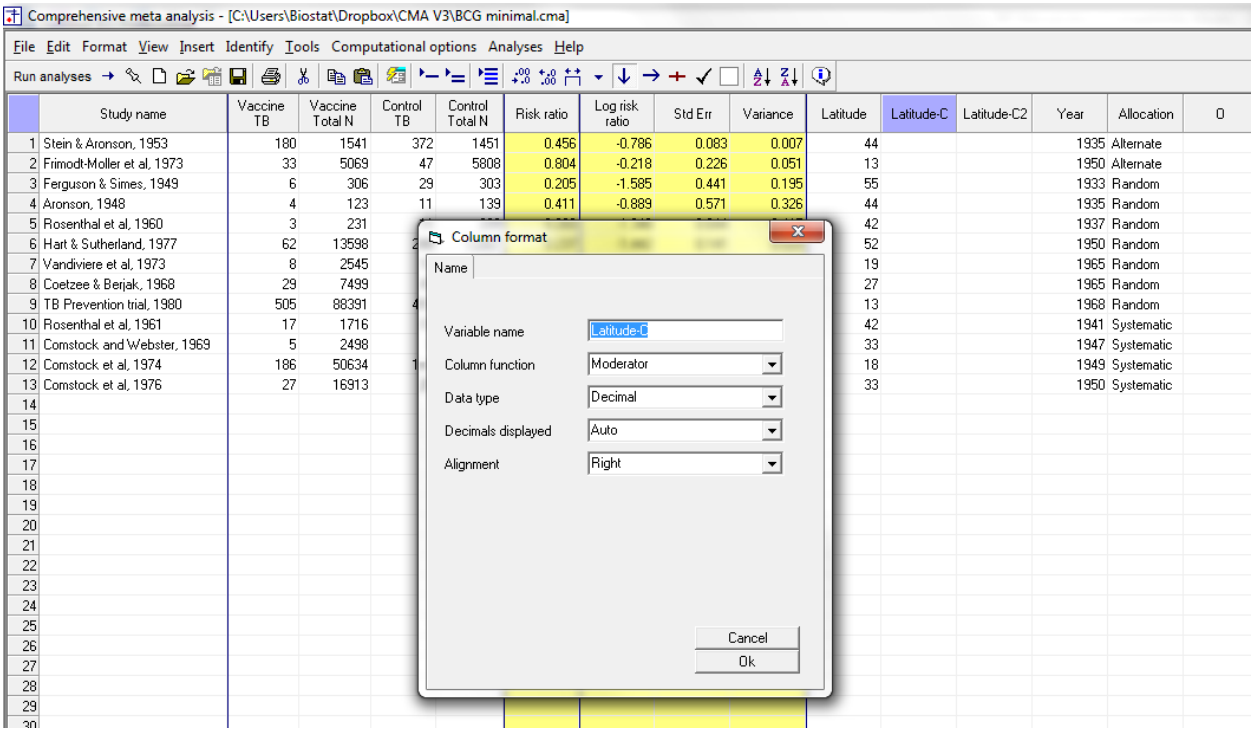
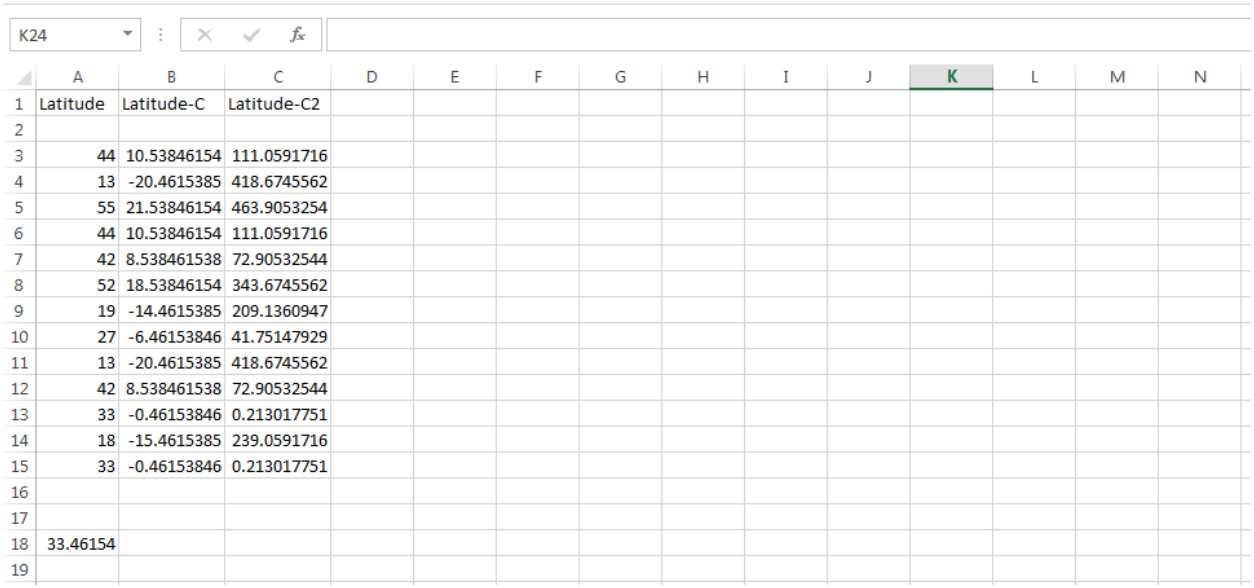


Figure 148 | Creating variables for interactions

- Click on the Dose Column
- Click Edit > Copy with Header

Open Excel™

- Paste the column into Column A
- Define Cell A18 as =AVERAGE(A3:A15)
- Define Cell B3 as =A3-A18 and copy to other rows
- Define Cell C3 as =B3^2 and copy to other rows
- Copy columns B and C (rows 3 to 15) to the clipboard

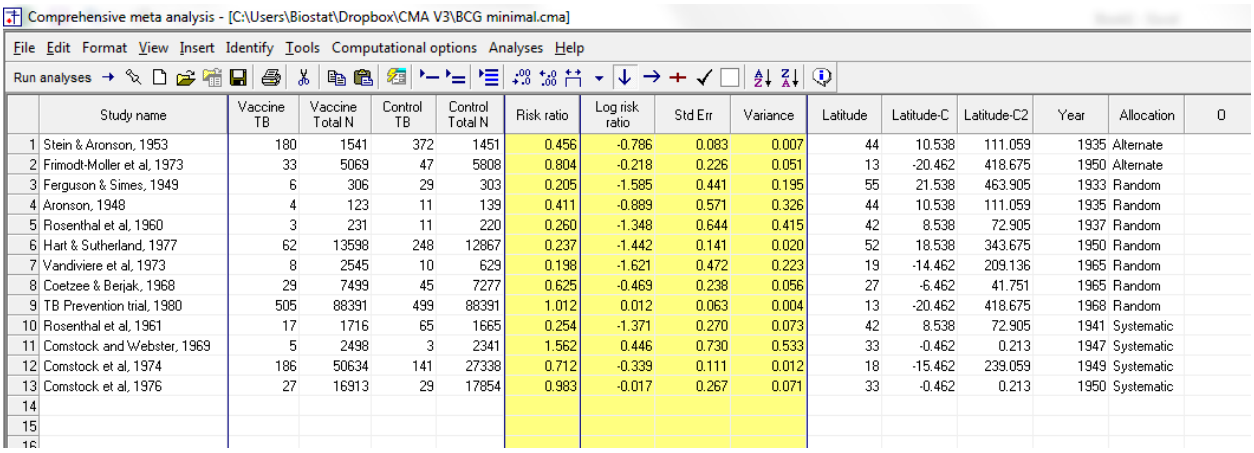


	A	B	C
1	Latitude	Latitude-C	Latitude-C2
2			
3	44	10.53846154	111.0591716
4	13	-20.4615385	418.6745562
5	55	21.53846154	463.9053254
6	44	10.53846154	111.0591716
7	42	8.538461538	72.90532544
8	52	18.53846154	343.6745562
9	19	-14.4615385	209.1360947
10	27	-6.46153846	41.75147929
11	13	-20.4615385	418.6745562
12	42	8.538461538	72.90532544
13	33	-0.46153846	0.213017751
14	18	-15.4615385	239.0591716
15	33	-0.46153846	0.213017751
16			
17			
18	33.46154		
19			

Figure 149 | Creating variables for interactions

Return to CMA

- Click on Row-1 in the Dose-C column
- Click CTRL-V to paste the data
- Save the file with the new data



Comprehensive meta analysis - [C:\Users\Biostat\Dropbox\CMA V3\BCG minimal.cma]

File Edit Format View Insert Identify Tools Computational options Analyses Help

Run analyses

	Study name	Vaccine TB	Vaccine Total N	Control TB	Control Total N	Risk ratio	Log risk ratio	Std Err	Variance	Latitude	Latitude-C	Latitude-C2	Year	Allocation	0
1	Stein & Aronson, 1953	180	1541	372	1451	0.456	-0.786	0.083	0.007	44	10.538	111.059	1935	Alternate	
2	Frimodt-Moller et al, 1973	33	5069	47	5808	0.804	-0.218	0.226	0.051	13	-20.462	418.675	1950	Alternate	
3	Ferguson & Simes, 1949	6	306	29	303	0.205	-1.585	0.441	0.195	55	21.538	463.905	1933	Random	
4	Aronson, 1948	4	123	11	139	0.411	-0.889	0.571	0.326	44	10.538	111.059	1935	Random	
5	Rosenthal et al, 1960	3	231	11	220	0.260	-1.348	0.644	0.415	42	8.538	72.905	1937	Random	
6	Hart & Sutherland, 1977	62	13598	248	12867	0.237	-1.442	0.141	0.020	52	18.538	343.675	1950	Random	
7	Vandiviere et al, 1973	8	2545	10	629	0.198	-1.621	0.472	0.223	19	-14.462	209.136	1965	Random	
8	Coetzee & Berjak, 1968	29	7499	45	7277	0.625	-0.469	0.238	0.056	27	-6.462	41.751	1965	Random	
9	TB Prevention trial, 1980	505	88391	499	88391	1.012	0.012	0.063	0.004	13	-20.462	418.675	1968	Random	
10	Rosenthal et al, 1961	17	1716	65	1665	0.254	-1.371	0.270	0.073	42	8.538	72.905	1941	Systematic	
11	Comstock and Webster, 1969	5	2498	3	2341	1.562	0.446	0.730	0.533	33	-0.462	0.213	1947	Systematic	
12	Comstock et al, 1974	186	50634	141	27338	0.712	-0.339	0.111	0.012	18	-15.462	239.059	1949	Systematic	
13	Comstock et al, 1976	27	16913	29	17854	0.983	-0.017	0.267	0.071	33	-0.462	0.213	1950	Systematic	
14															
15															

Figure 150 | Creating variables for interactions

APPENDIX 7: PLOTTING INTERACTIONS

In this appendix we show how to plot three kinds of interactions

- The interaction of two categorical covariates
- The interaction of one categorical and one continuous covariate
- The interaction of two continuous covariates

All three follow basically the same format in Excel™.

The spreadsheet that we use in this appendix can be downloaded at _____. Be sure to select the correct tab at the bottom of the spread

Plotting the interaction of two categorical covariates

In chapter we ran an analysis to assess the interaction of High (I) by Long (I). The results of the analysis are shown in Figure 151. To create the plot we copy the covariate names and coefficients from CMA into Excel™ as shown in Figure 152.

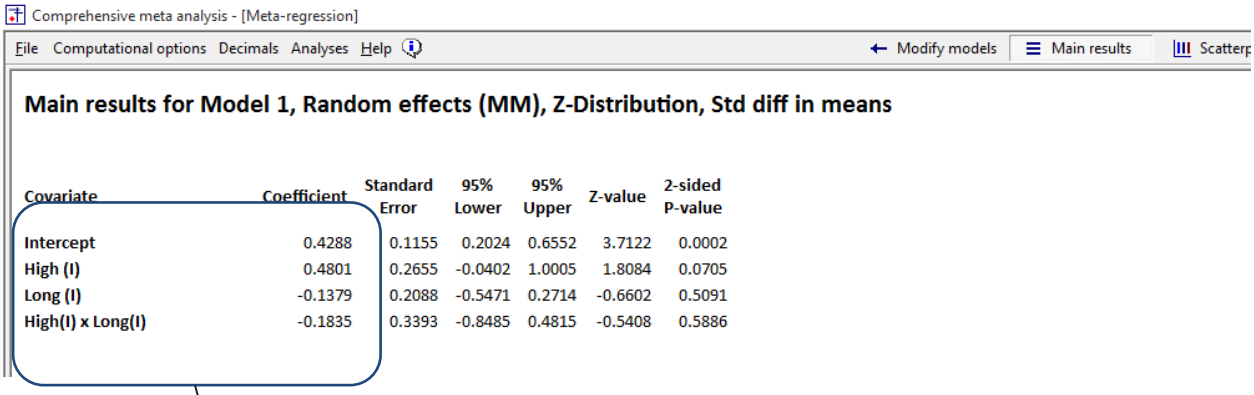


Figure 151 | Plotting interaction of two categorical covariates

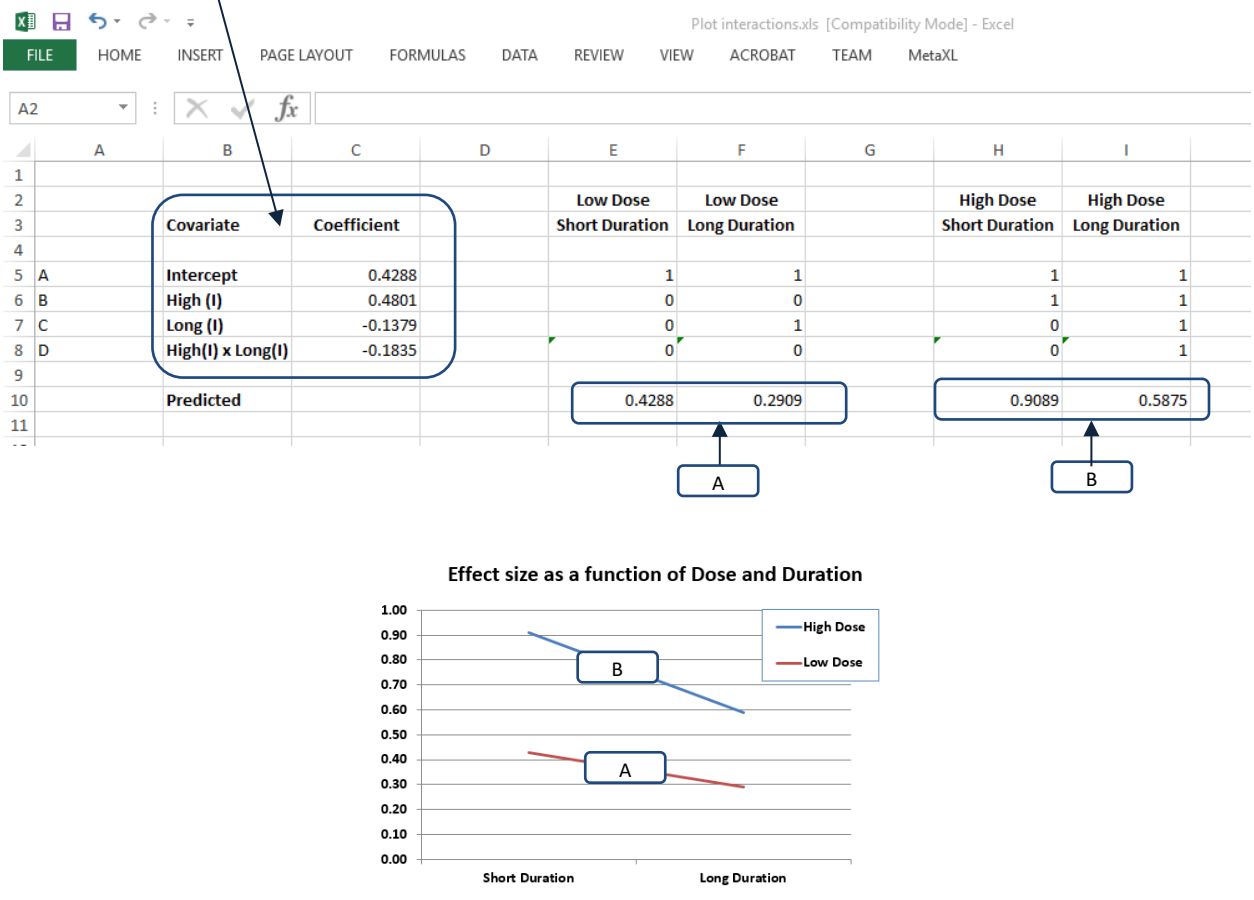


Figure 152 | Plotting interaction of two categorical covariates

Consider the column called “Low Dose Short Duration”

- The value for Intercept is 1
- The value for High (I) is 0 (since this study is coded 0 for High-Dose)
- The value for Long (I) is 0 (since this study is coded 0 for Long Duration)
- The value for High (I) x Long (I) is 0 (this is the product of 0 times 0)

To get the predicted value we multiply each X value in column E by the corresponding coefficient in column C, and the sum across the four rows. The predicted effect for this column is then

$$Y = 0.4288 \times 1 + 0.4801 \times 0 - 0.1379 \times 0 - 0.1835 \times 0 = 0.4288 \qquad (1.119)$$

Using the same logic we compute the predicted effect for four points.

To create the plot we need to identify the cells E10 and F10 as endpoints for the Low-Dose studies, and H10 and I10 as endpoints for the High-Dose studies.

The specific steps vary depending on the version of Excel. The instructions below are for Excel 2013.

Insert > Line Chart > 2D Line

Design > Select data > Add Series

Series name	H2	High Dose
Series Y-Values	H10:I10	0.9089, 0.5875

Design > Select data > Add Series

Series name	E2	Low Dose
Series Y-Values	E10:F10	0.4288, 0.2909

Design > Select data > Horizontal Axis Labels

Edit	H3:I3	Short, Long

Design > Add chart element > Chart Title > Effect size as a function of Dose and Duration

Design > Add chart element > Legend

Plotting the interaction of a categorical covariate by a continuous covariate

In chapter we ran an analysis to assess the interaction of High (I) by Long (I). The results of the analysis are shown in Figure 151. To create the plot we'll need the covariate names and coefficients from this figure. Copy these to Excel™ columns B and C as shown in Figure 152.

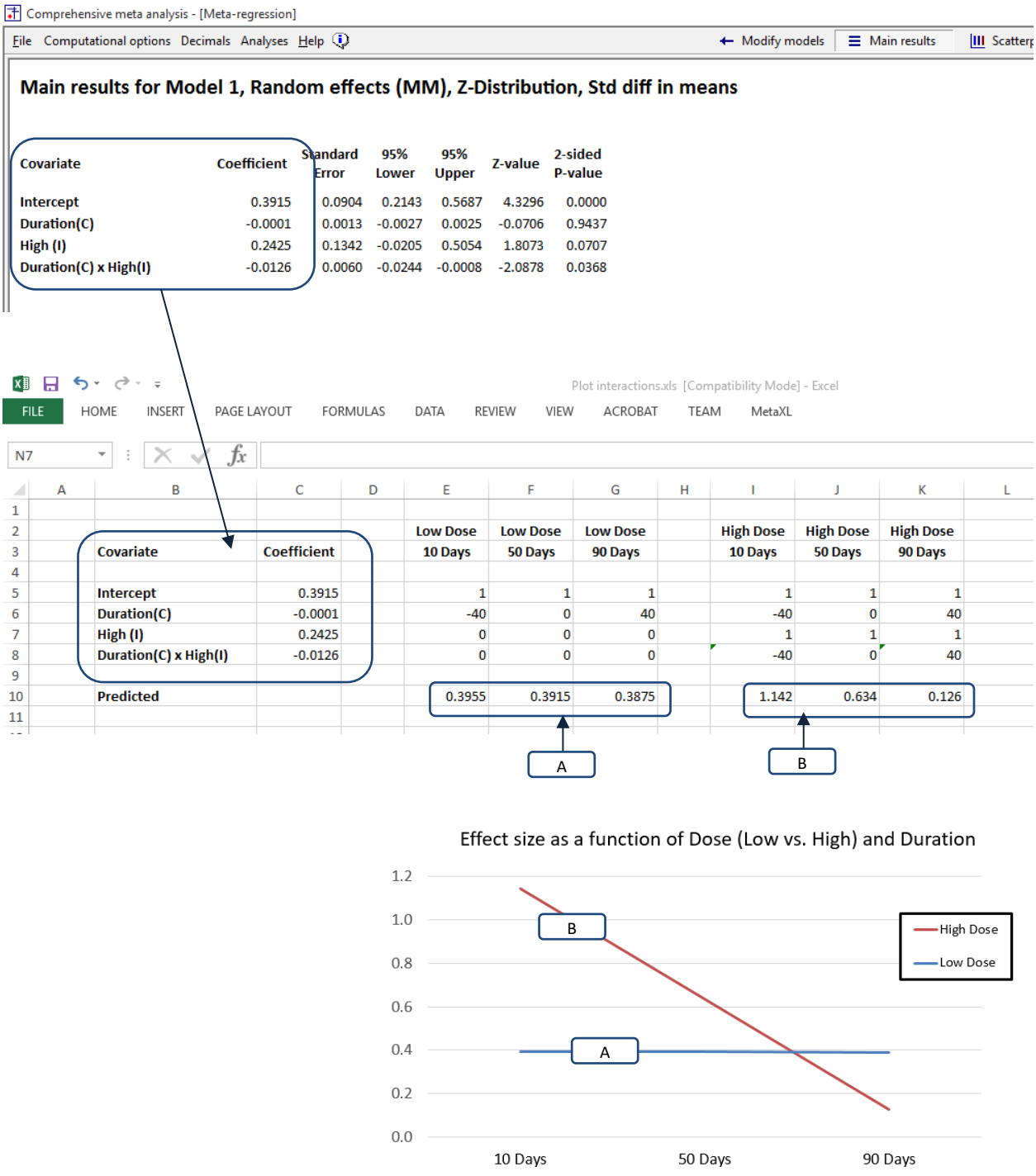


Figure 153 | Plotting interaction of two categorical covariates

Consider the column called “Low Dose 10 Days”

- The value for Intercept is 1
- The Duration for this column is 10, but the covariate is Duration (C), which is Duration minus the mean (50). So, we enter 10 minus 50, or -40.
- The value for Long (I) is 0 (since this study is coded 0 for Long Duration)
- The value for Duration (C) x Long (I) is 0 (this is the product of -40 times 0)

To get the predicted value we multiply each X value in column E by the corresponding coefficient in column C, and the sum across the four rows. The predicted effect for this column is then

$$Y = 0.3915 \times 1 - 0.0001 \times -40 + 0.2424 \times 0 - 0.0126 \times 0 = 0.3955 \qquad (1.120)$$

Using the same logic we compute the predicted effect for four points.

Note that a 50-day study is coded 0 for Duration (C) while a 90-day study is coded 40. For the lines in the plot to be straight, we need the three data-points to be evenly spaced (that is, -40, 0, +40).

To create the plot we need to identify the cells I10 to K10 as data-points for the High-Dose studies, and E10 to G10 as data-points for the Low-Dose studies.

The specific steps vary depending on the version of Excel. The instructions below are for Excel 2013.

Insert > Line Chart > 2D Line

Design > Select data > Add Series

Series name	I2	High Dose
Series Y-Values	I10:K10	1.1420, 0.6340, 0.1260

Design > Select data > Add Series

Series name	E2	Low Dose
Series Y-Values	E10:G10	0.3955, 0.3915, 0.3875

Design > Select data > Horizontal Axis Labels

Edit	I3:K3	10 Days, 50 Days, 90 Days

Design > Add chart element > Chart Title > Effect size as a function of Dose and Duration

Design > Add chart element > Legend

Plotting the interaction of two continuous covariates

In chapter we ran an analysis to assess the interaction of High (I) by Long (I). The results of the analysis are shown in Figure 151. To create the plot we'll need the covariate names and coefficients from this figure. Copy these to Excel™ columns B and C as shown in Figure 152.

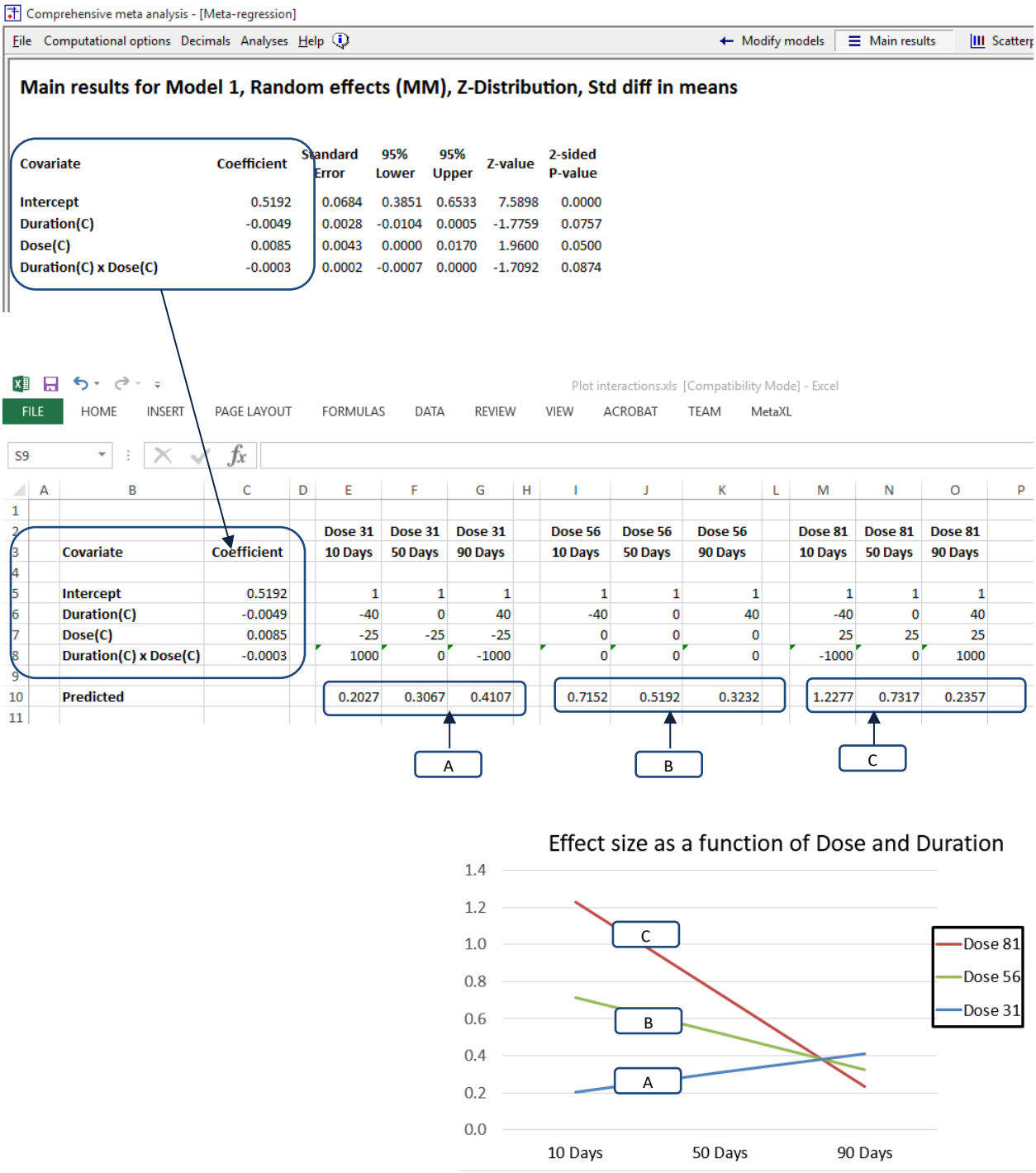


Figure 154 | Plotting interaction of two categorical covariates

Consider the column called “Low Dose 10 Days”

- The value for Intercept is 1
- The Duration for this column is 10, but the covariate is Duration (C), which is Duration minus the mean (50). So, we enter 10 minus 50, or -40.
- The Dose for this column is 31, but the covariate is Dose (C), which is Dose minus the mean (56). So, we enter 31 minus 56, or -25.
- The value for Duration (C) x Dose (C) is 1000 (this is the product of -40 times -25)

To get the predicted value we multiply each X value in column E by the corresponding coefficient in column C, and the sum across the four rows. The predicted effect for this column is then

$$Y = 0.5192 \times 1 - 0.0049 \times -40 + 0.0085 \times (-25) - 0.0126 \times 1000 = 0.2027 \qquad (1.121)$$

Using the same logic we compute the predicted effect for all points.

A 50-day study is coded 0 for Duration (C) while a 90-day study is coded 40. For the lines in the plot to be straight, we need the three data-points to be evenly spaced (that is, -40, 0, +40).

A dose of 56 is coded 0 for Dose (C) while a Dose of 81 is coded 25. For the lines in the plot to be equally spaced, we need the three data-points to be evenly spaced (that is, -25, 0, +25).

To create the plot we need to identify the cells M10 to O10 as data-points for the 81 mg studies, I10 to K10 as data-points for the 56 mg studies, and E10 to G10 for the 31 mg studies

The specific steps vary depending on the version of Excel. The instructions below are for Excel 2013.

Insert > Line Chart > 2D Line

Design > Select data > Add Series

Series name	M2	Dose 81
Series Y-Values	M10:O10	1.2277, 0.7317, 0.2357

Design > Select data > Add Series

Series name	I2	Dose 51
Series Y-Values	I10:K10	0.7152, 0.5192, 0.3232

Design > Select data > Add Series

Series name	E2	Dose 31
Series Y-Values	E10:G10	0.2027, 0.3067, 0.4107

Design > Select data > Horizontal Axis Labels

Edit	M3:O3	10 Days, 50 Days, 90 Days

Design > Add chart element > Chart Title > Effect size as a function of Dose and Duration

Design > Add chart element > Legend

APPENDIX 6: PLOTTING A CURVILINEAR RELATIONSHIP

In chapter Interaction we showed how to use Dose and Dose^2 to predict effect size. The program will not plot a curvilinear relationship, so we need to create the plot in Excel™.

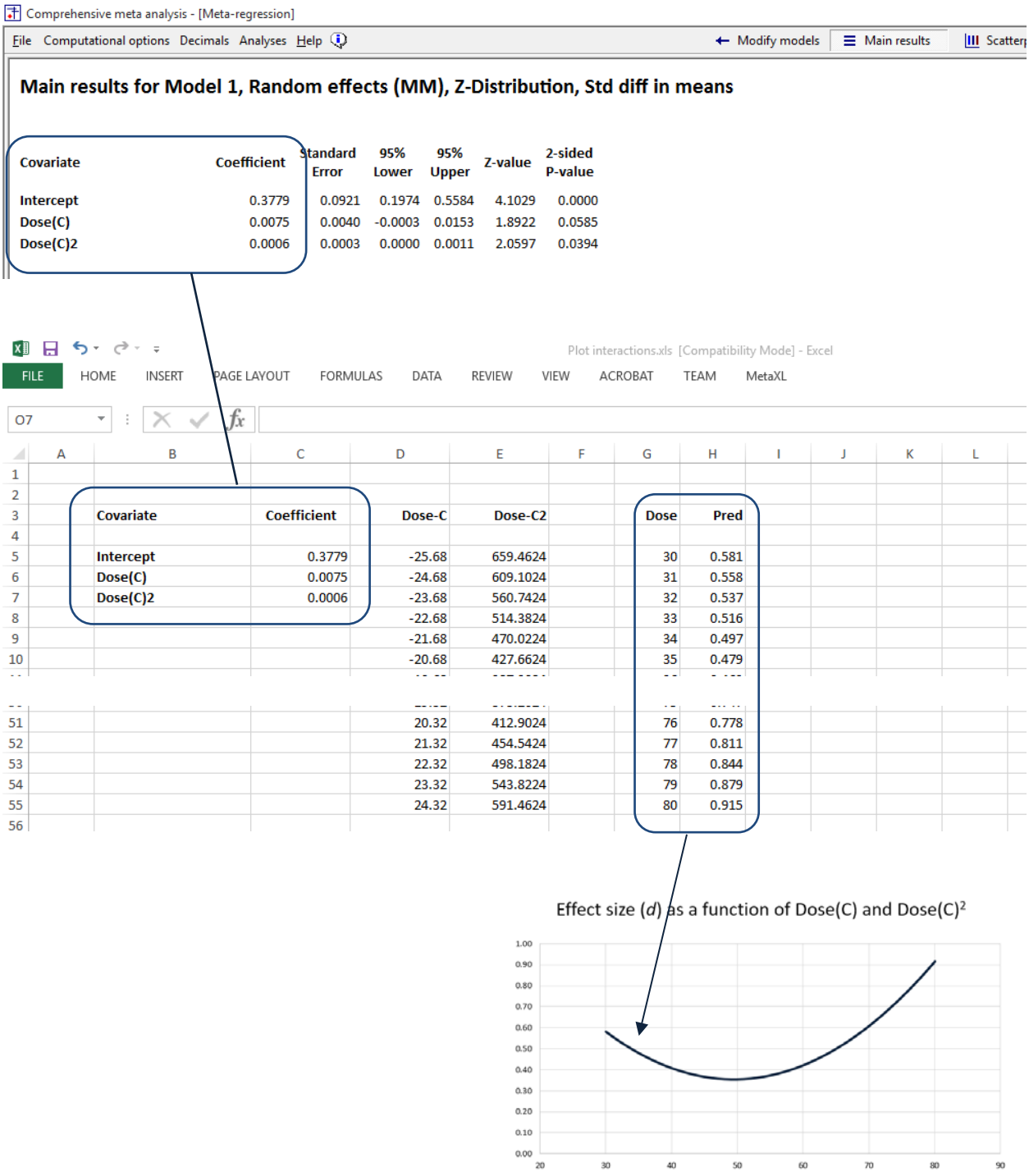


Figure 155 | Plotting a curvilinear relationship

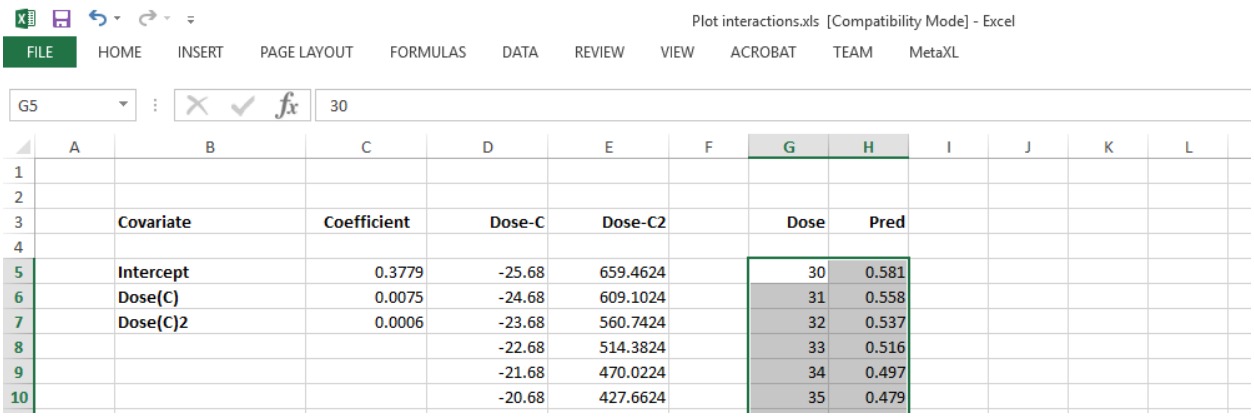
Copy the covariate names into Column B
Copy the coefficients into Column C

Column B	Covariates	Copy names from output
Column C	Coefficients	Copy names from output
Column G	Dose	Enter 30 in Row 5, then add 1 for each subsequent row
Column D	Dose(C)	Dose minus 55.68
Column E	Dose(C)	Dose(C) squared
Column H	Predicted	= 0.3379 + D(0.0075) + E (0.0006)

At this point column G holds the doses while Column H holds the predicted values. We can use these two columns to create a plot.

The instructions here are for Microsoft Excel™ 2013. The specific commands may vary slightly for other versions.

- Highlight columns G and H from row 3 to the bottom (Figure 156)
- Select Insert > Scatter > Scatter with smooth lines
- The program creates the plot shown above



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3		**Covariate**	**Coefficient**	**Dose-C**	**Dose-C2**		**Dose**	**Pred**				
4												
5		**Intercept**	0.3779	-25.68	659.4624		30	0.581				
6		**Dose(C)**	0.0075	-24.68	609.1024		31	0.558				
7		**Dose(C)2**	0.0006	-23.68	560.7424		32	0.537				
8				-22.68	514.3824		33	0.516				
9				-21.68	470.0224		34	0.497				
10				-20.68	427.6624		35	0.479				

Figure 156 | Plotting a curvilinear relationship

APPENDIX 9: META-REGRESSION IN STATA

In this section we show the correspondence between results produced by CMA and those produced by the stata macro “metareg”

ADD PLOTTING RATIOS, PREVALENCE

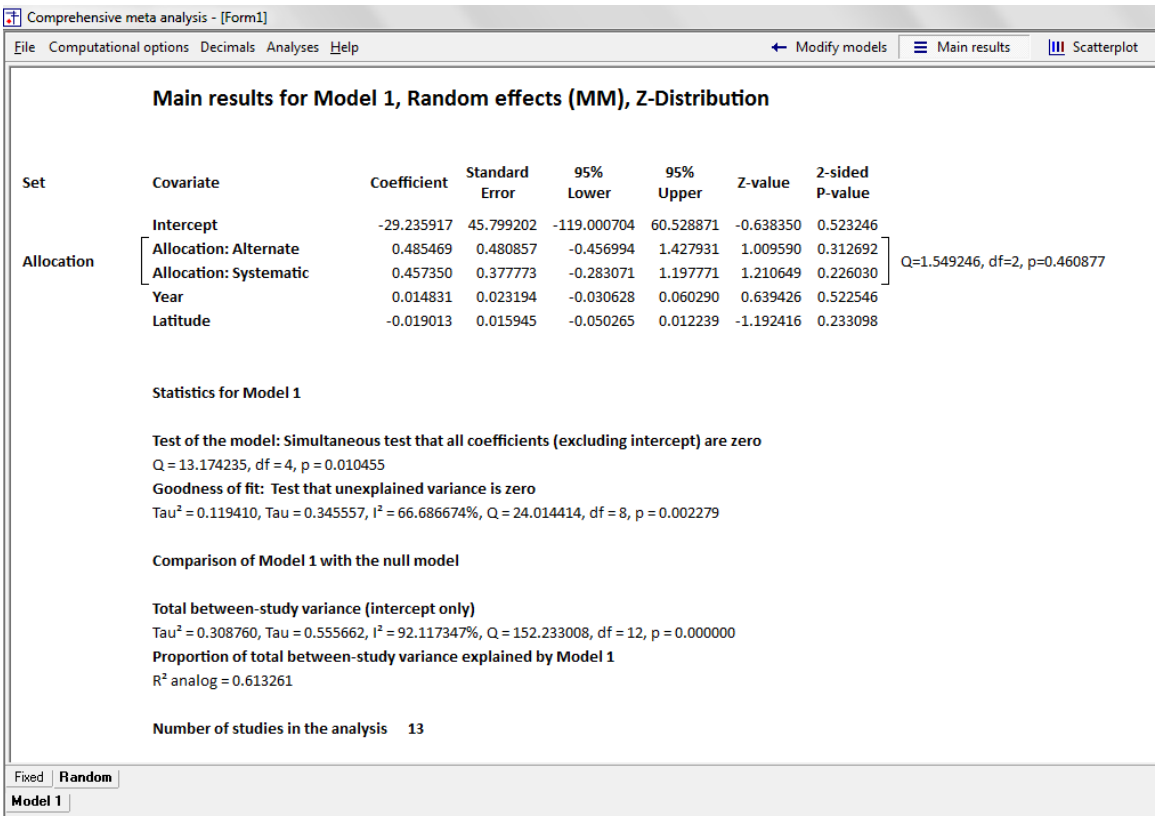


Figure 157 | CMA | Intercept + Year + Dose + Allocation | Z | Method of moments

```
. metareg logrr alternate systematic year lat, wsse(se) bsest(mm)  tau2test z

Meta-regression                                   Number of obs  =      13
Method of moments estimate of between-study variance  tau2       =   .1194
% residual variation due to heterogeneity         I-squared_res  =  66.69%
Proportion of between-study variance explained    Adj R-squared  =  61.33%
Joint test for all covariates                     Model chi2(4)  =   13.17
Without Knapp-Hartung modification                Prob > chi2    =  0.0105

------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   alternate |   .4854685   .4808573     1.01   0.313    -.4569944    1.427931
  systematic |     .45735   .3777728     1.21   0.226     -.283071    1.197771
        year |   .0148307   .0231938     0.64   0.523    -.0306283    .0602897
         lat |  -.0190132   .0159451    -1.19   0.233    -.0502649    .0122386
       _cons |  -29.23591    45.7992    -0.64   0.523    -119.0007    60.52887
------------------------------------------------------------------------------
Test for residual between-study variance (of tau2=0)  Q_res (8 df)   =   24.01
                                                      Prob > Q_res   =  0.0023
```

Figure 158 | Metareg| Intercept + Year + Dose + Allocation | *Z* | Method of moments

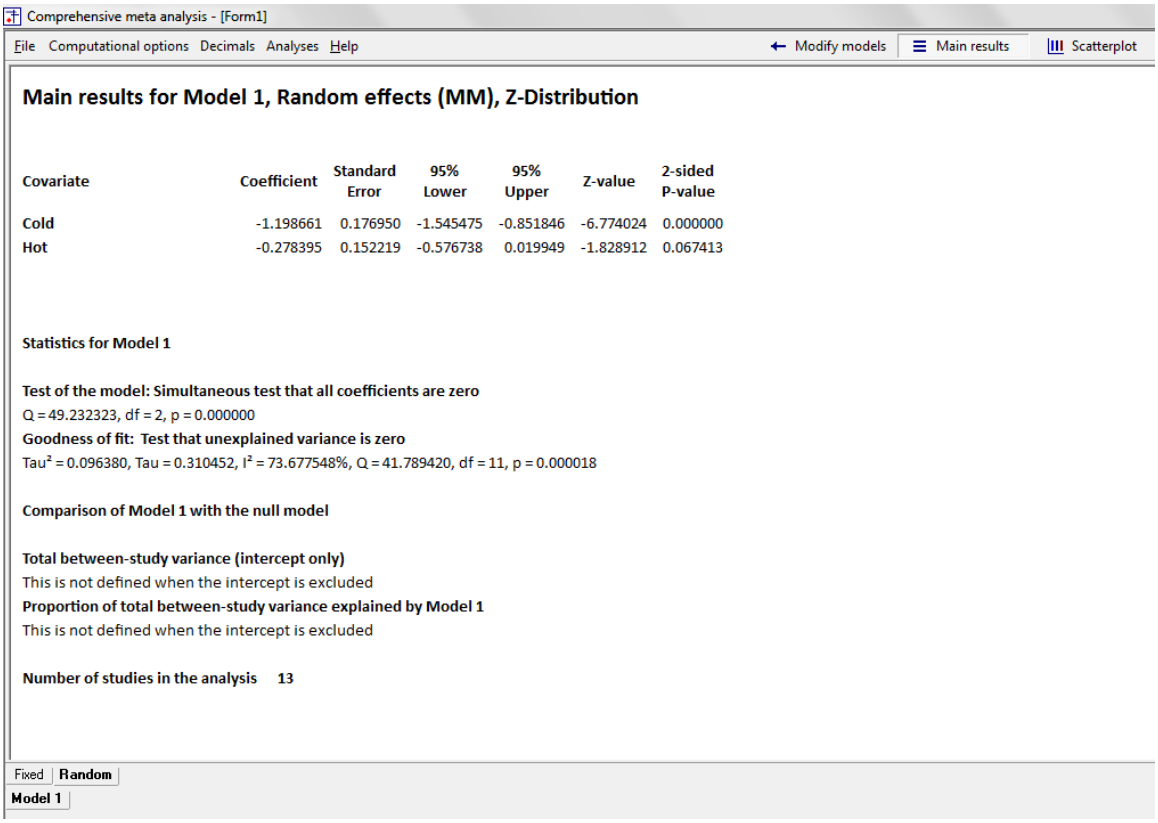


Figure 159 | CMA | Allocation | Z | Method of moments

```
. metareg logrr cold hot, wsse(se) bsest(mm)  tau2test z noconst

Meta-regression                                     Number of obs  =      13
Method of moments estimate of between-study variance  tau2         =  .09638
% residual variation due to heterogeneity           I-squared_res  =  73.68%
Joint test for all covariates                       Model chi2(2)  =   49.23
Without Knapp-Hartung modification                  Prob > chi2    =  0.0000

------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cold |  -1.198661   .1769496    -6.77   0.000    -1.545475   -.8518459
         hot |   -.278395   .1522189    -1.83   0.067    -.5767385    .0199485
------------------------------------------------------------------------------
Test for residual between-study variance (of tau2=0)  Q_res (11 df)  =   41.79
                                                      Prob > Q_res   =  0.0000
```

Figure 160 | Metareg | Allocation | *Z* | Method of moments

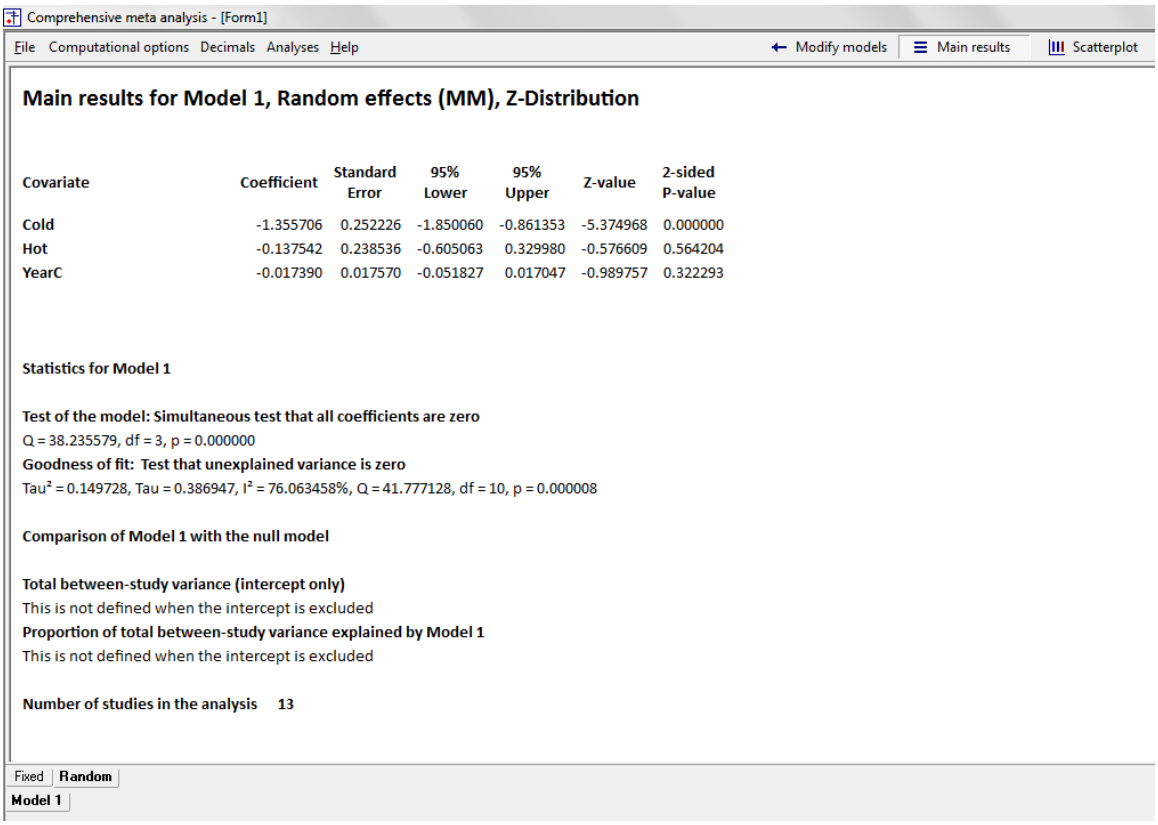


Figure 161 | CMA | Allocation, Year | *Z* | Method of moments

```
. metareg logrr cold hot yearc, wsse(se) bsest(mm)  tau2test z noconst

Meta-regression                                     Number of obs  =      13
Method of moments estimate of between-study variance  tau2         =   .1497
% residual variation due to heterogeneity           I-squared_res  =  76.06%
Joint test for all covariates                       Model chi2(3)  =   38.24
Without Knapp-Hartung modification                  Prob > chi2    =  0.0000

------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cold |  -1.355706   .2522259    -5.37   0.000    -1.85006   -.8613525
         hot |  -.1375418   .2385357    -0.58   0.564    -.6050632    .3299796
       yearc |  -.0173902   .0175702    -0.99   0.322    -.0518272    .0170467
------------------------------------------------------------------------------
Test for residual between-study variance (of tau2=0)  Q_res (10 df)  =   41.78
                                                      Prob > Q_res   =  0.0000
```

Figure 162 | Metareg | Allocation, Year | *Z* | Method of moments

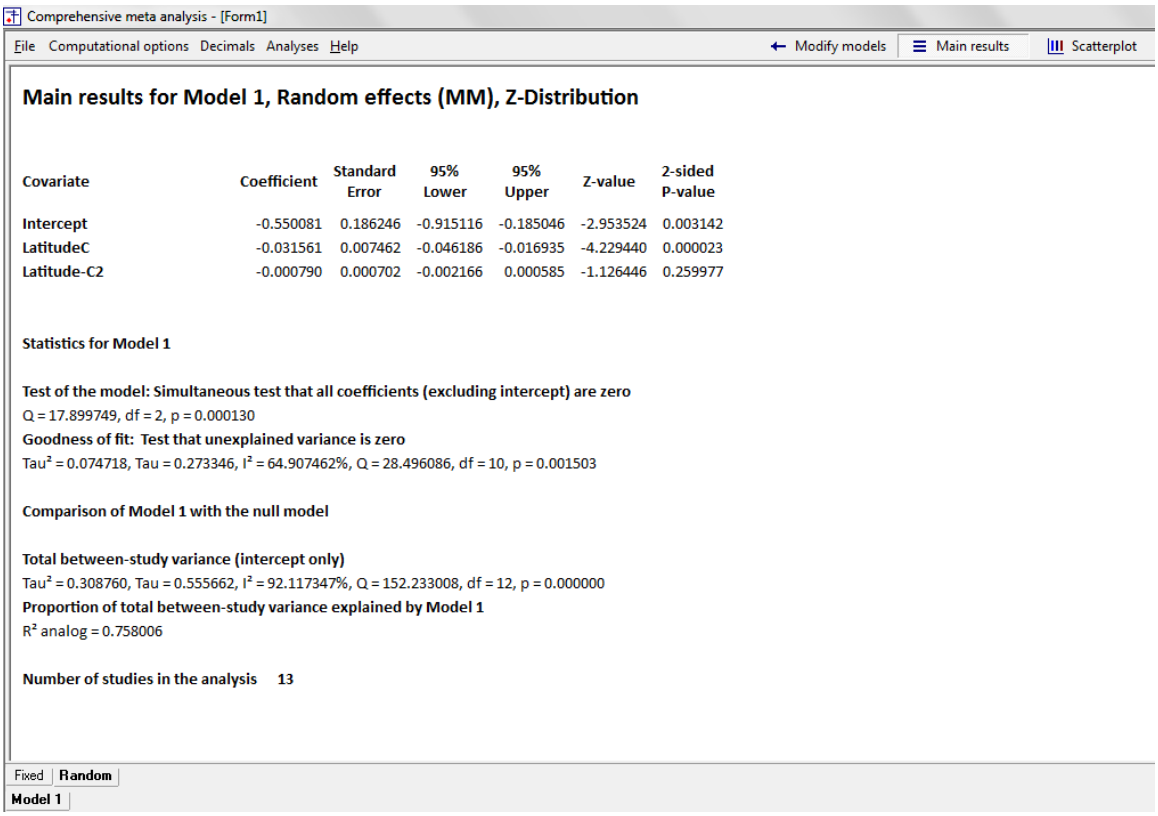


Figure 163 | CMA | Intercept, Year-C, Year-C2 | *Z* | Method of moments

```
. metareg logrr latc latc2, wsse(se) bsest(mm)  tau2test z

Meta-regression                                     Number of obs  =      13
Method of moments estimate of between-study variance  tau2         =  .07472
% residual variation due to heterogeneity           I-squared_res  =  64.91%
Proportion of between-study variance explained      Adj R-squared  =  75.80%
Joint test for all covariates                       Model chi2(2)  =   17.90
Without Knapp-Hartung modification                  Prob > chi2    =  0.0001

------------------------------------------------------------------------------
       logrr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        latc |  -.0315607   .0074622    -4.23   0.000    -.0461863   -.0169352
       latc2 |  -.0007904   .0007017    -1.13   0.260    -.0021656    .0005849
       _cons |  -.5500812   .1862457    -2.95   0.003    -.9151161   -.1850463
------------------------------------------------------------------------------
Test for residual between-study variance (of tau2=0)  Q_res (10 df)  =   28.50
                                                      Prob > Q_res   =  0.0015
```

Figure 164 | Metareg | Intercept, Year-C, Year-C2 | *Z* | Method of moments

REFERENCES

Berkey, C.S., Hoaglin, D.C. Mosteller, F., and Colditz, G.A. (1995) A random-effects regression model for meta-analysis. *Statistics in Medicine*, 14, 395-411.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H. (2009) *Introduction to Meta-Analysis*. Chichester: Wiley.

Cohen J., Cohen P., West S.G., Aiken, L.S. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences 3rd Edition*. Mahwah: Lawrence Erlbaum Associates

Colditz, G.A., Brewer, F.B., Berkey, C.S., Wilson, E.M., Burdick, E., Gineberg, H.V., and Mosteller, F. (1994). Efficacy of BGC vaccine in the prevention of tuberculosis. *Journal of the American Medical Association* 271, 698-702.

Egger. M, Smith, G.W., Altman, D.G. (2001) *Systematic Reviews in Health Care: Meta-Analysis in Context*. (2nd Edition) London: BMJ Books

Hartung, J., Knapp, G., Sinha, B.K., (2008) *Statistical Meta-Analysis with Applications*. Hoboken: Wiley.

Hedges, L.V., and Olkin I. (1985) *Statistical Methods for Meta-Analysis*. Boston: Academic Press.

Hedges, L. and Pigott, T.D. (2001) The power of statistical tests in meta-analysis. *Psychological Methods* 6, 203-217.

Hedges, L. and Pigott, T.D. (2004) The power of statistical tests for moderators in meta-analysis. *Psychological Methods* 9, 426-445.

Higgins, J.P.T., and Thompson, S.G. (2004) Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine*, 23, 1663-1682.

Rothman, K.J. (1990). No Adjustments are Needed for Multiple Comparisons, *Epidemiology*, 1, 43-46.

Sutton, A.J., Abrams, K.R., Jones, D.R., Song, F. (2000) *Methods for Meta-Analysis in Medical Research*. Chichester: John Wiley and Sons

STEP 1: ENTER THE DATA

Insert column for study names

In Figure 165 [A], click Insert > Column for > Study names.

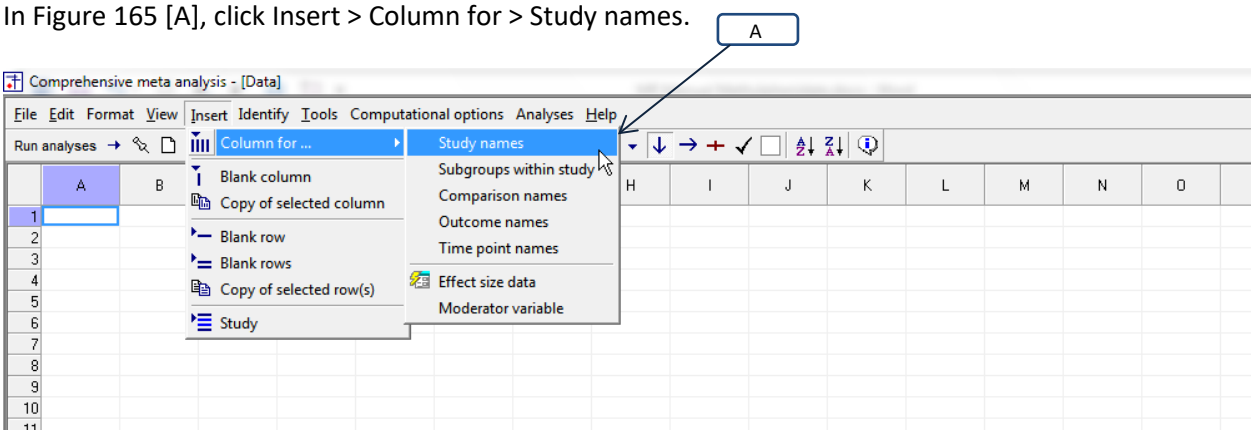


Figure 165 | Data-entry | Step 01

Figure 166 [B], the program creates a column labeled “Study name”.

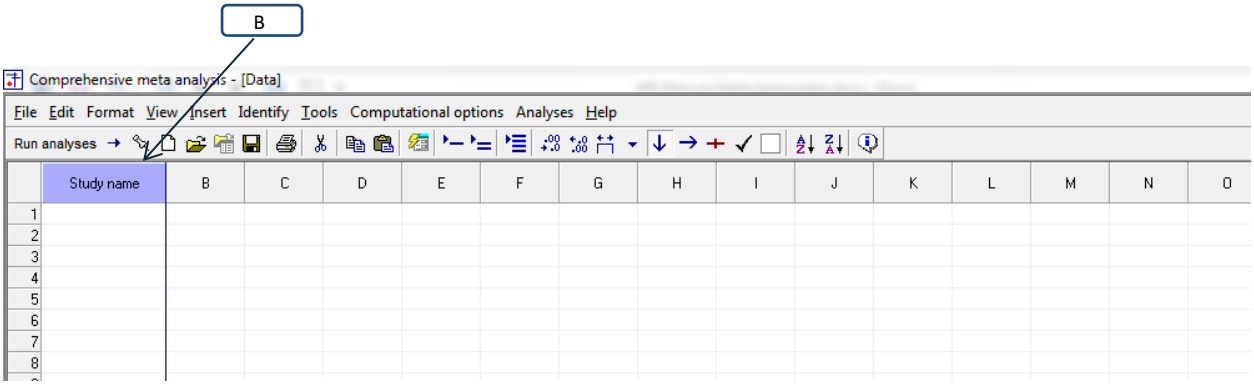


Figure 166 | Data-entry | Step 02

Insert columns for effect size data

In Figure 167, click Insert > Column for > Effect size data.

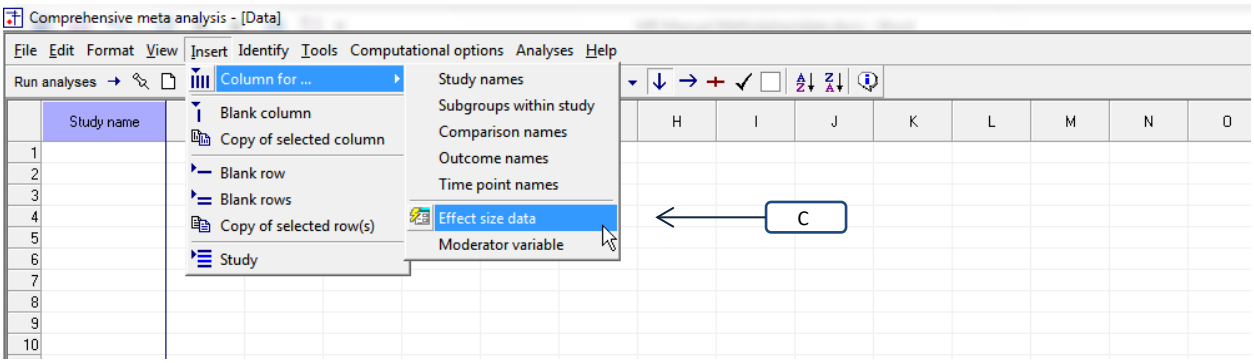


Figure 167 | Data-entry | Step 03

The program opens a wizard (Figure 168) that allows you to specify the kind of summary data you will enter

- Select <Show all 100 formats> [D]
- Click [Next] [E]

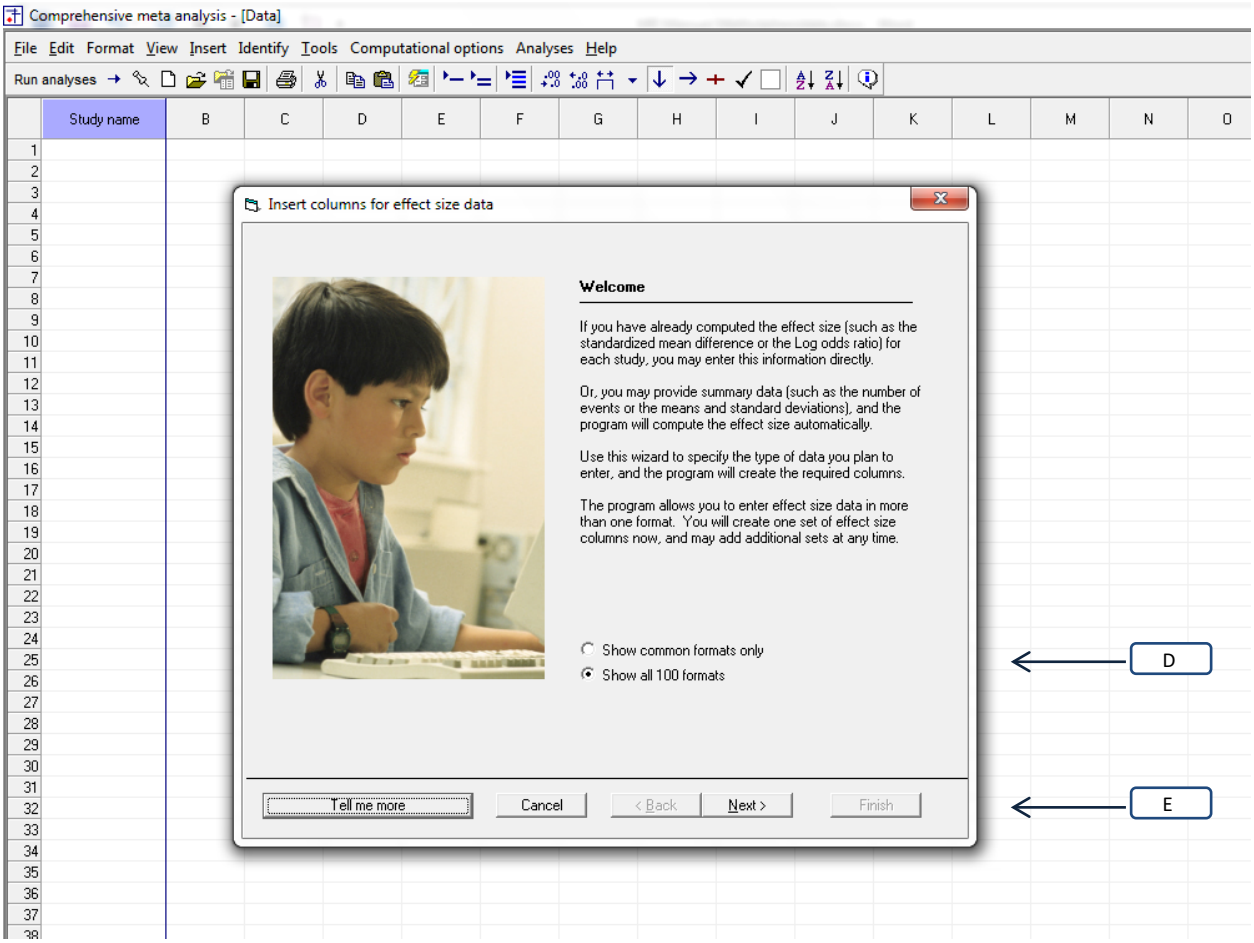


Figure 168 | Data-entry | Step 04

In the wizard (Figure 169)

- Select the top option button [F]
- On this screen, Click [Next] [G]

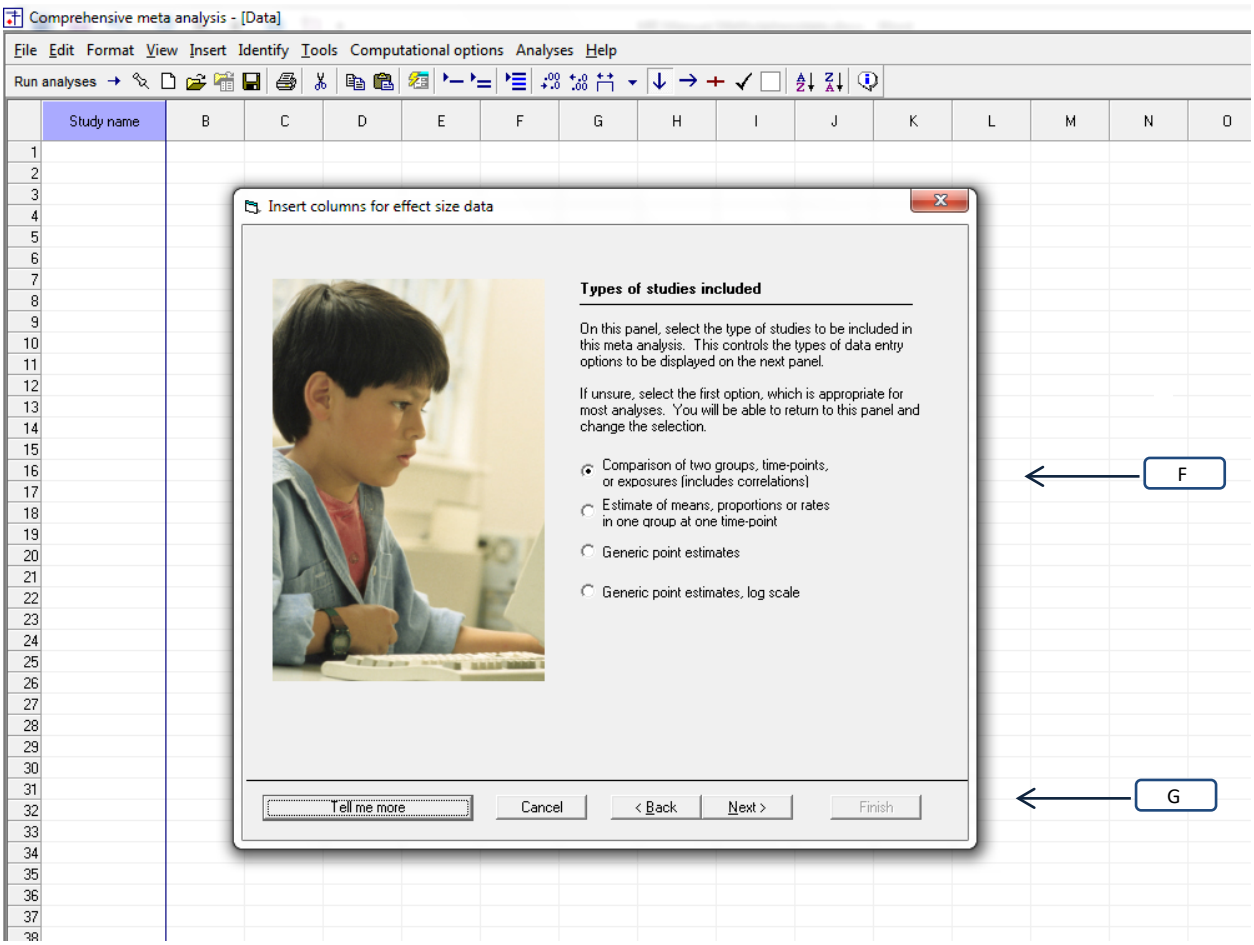


Figure 169 | Data-entry | Step 05

Typically in CMA we would enter summary data such as means and standard deviations, or events and sample size, and the program would compute the effect size for each study. In this example, however, the publication reported the effect size (*d*) and standard error for each study, and so we will enter that information directly.

In Figure 170, drill down to

- Continuous (means)
- Computed effect sizes
- Cohen's *d* and standard error [H]

Then, click <Finish>

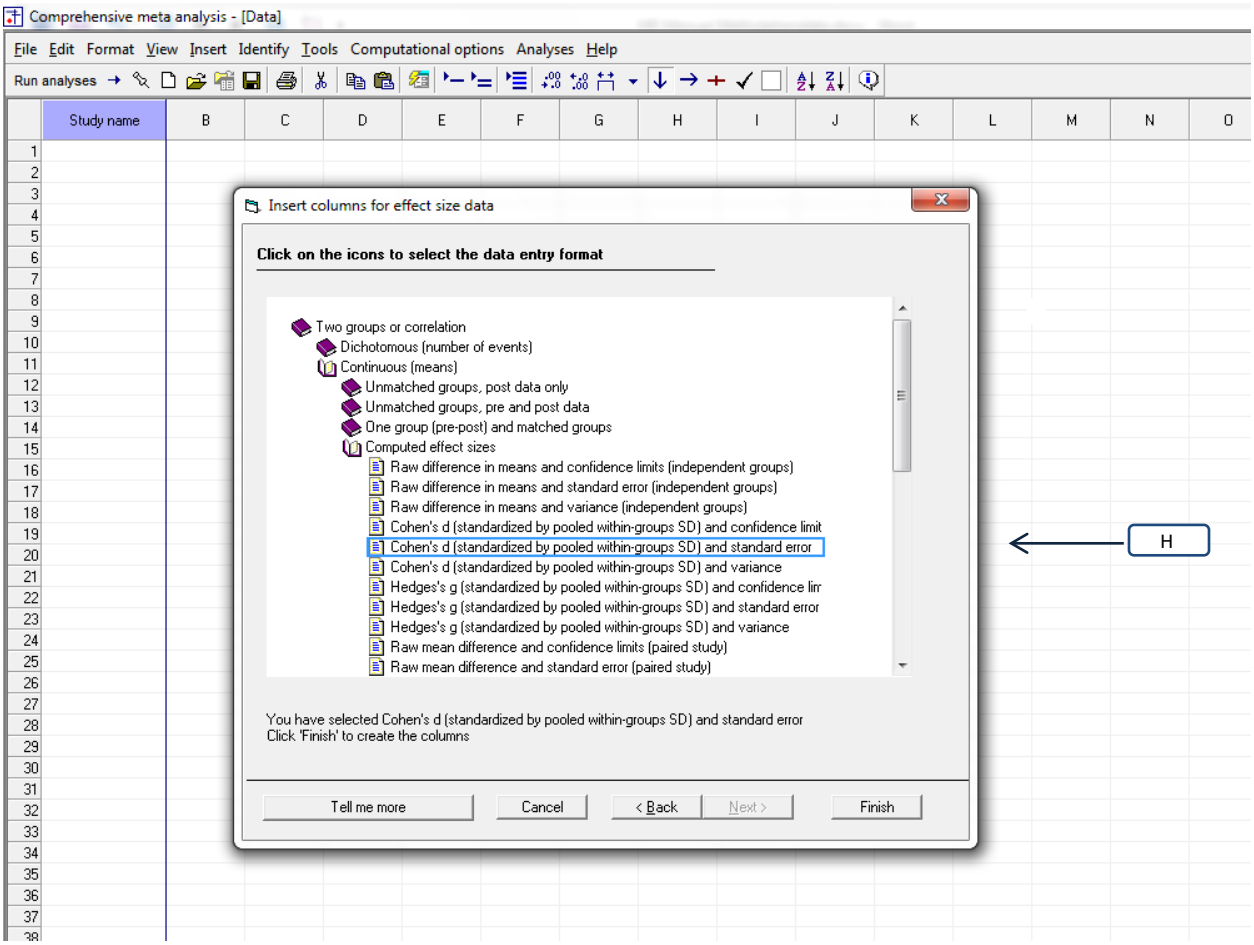


Figure 170 | Data-entry | Step 06

The program creates columns as shown in Figure 171. It also opens a wizard that allows you to label the columns.

- Enter Treated/Control as names for the two groups [I]

Then, click [Ok]

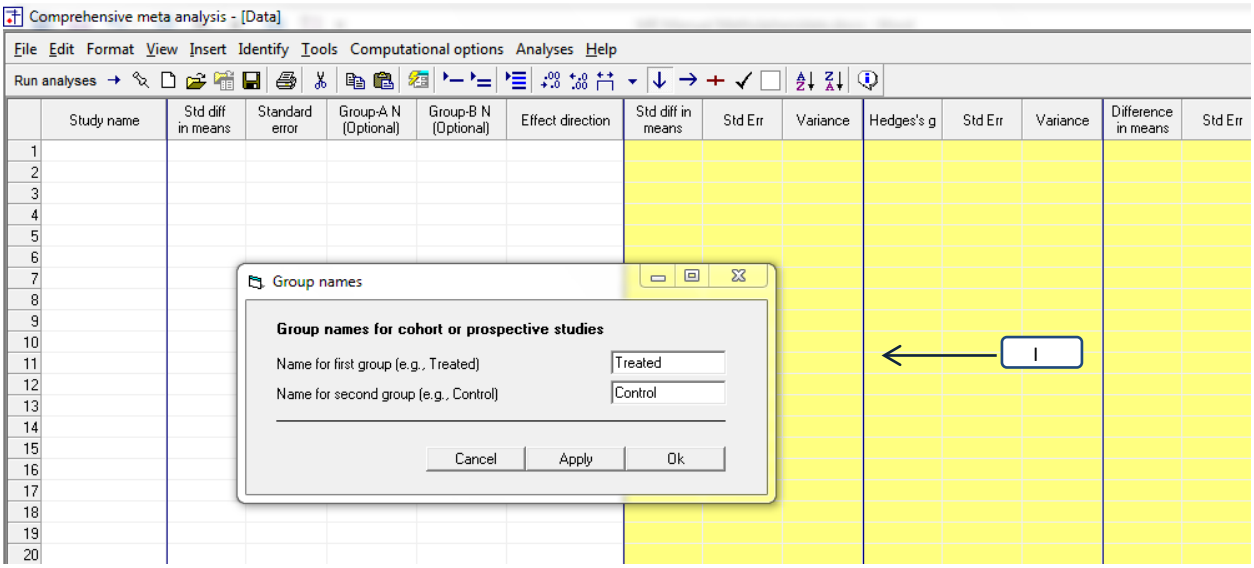


Figure 171 | Data-entry | Step 07

The program applies the labels as shown in Figure 172 [J].

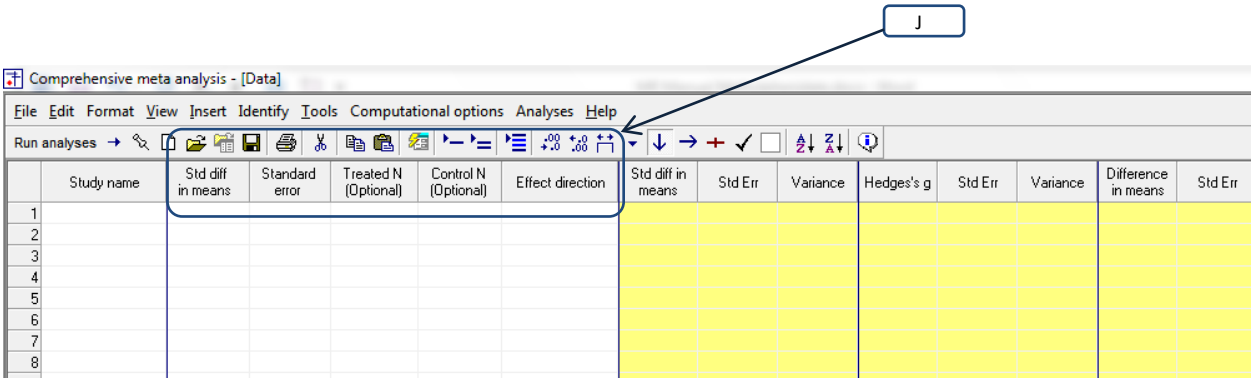


Figure 172 | Data-entry | Step 08

Customize the screen

The program initially displays three effect sizes – Standard difference in means (Cohen's *d*), Hedges's *g*, and the raw mean difference (Figure 173).

- We want to work with only Standard difference in means (Cohen's *d*).

Therefore customize the display as follows.

- Right-click in any yellow column
- Click <Customize computed effect size display> [A]

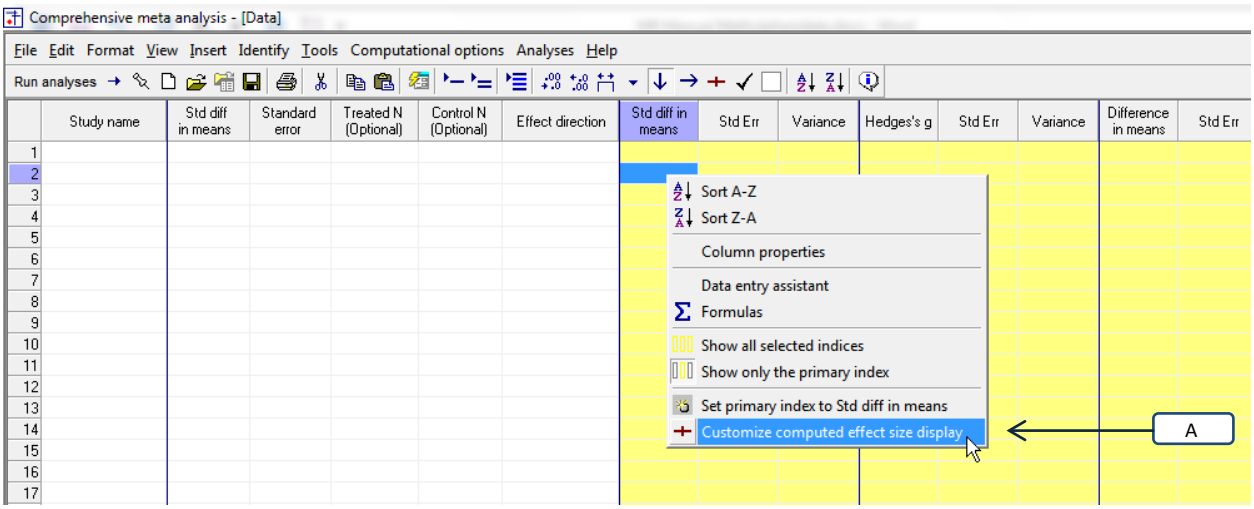


Figure 173 | Data-entry | Step 14

The program displays this wizard (Figure 174)

- Select the box for Std diff in means [B]
- De-select the box for Hedges's *g* [B]
- De-select the box for Difference in means [B]

- Select the box for Also show standard error [C]
- Select the box for Also show variance [C]

- Select "Std diff in means" as the primary index [D]

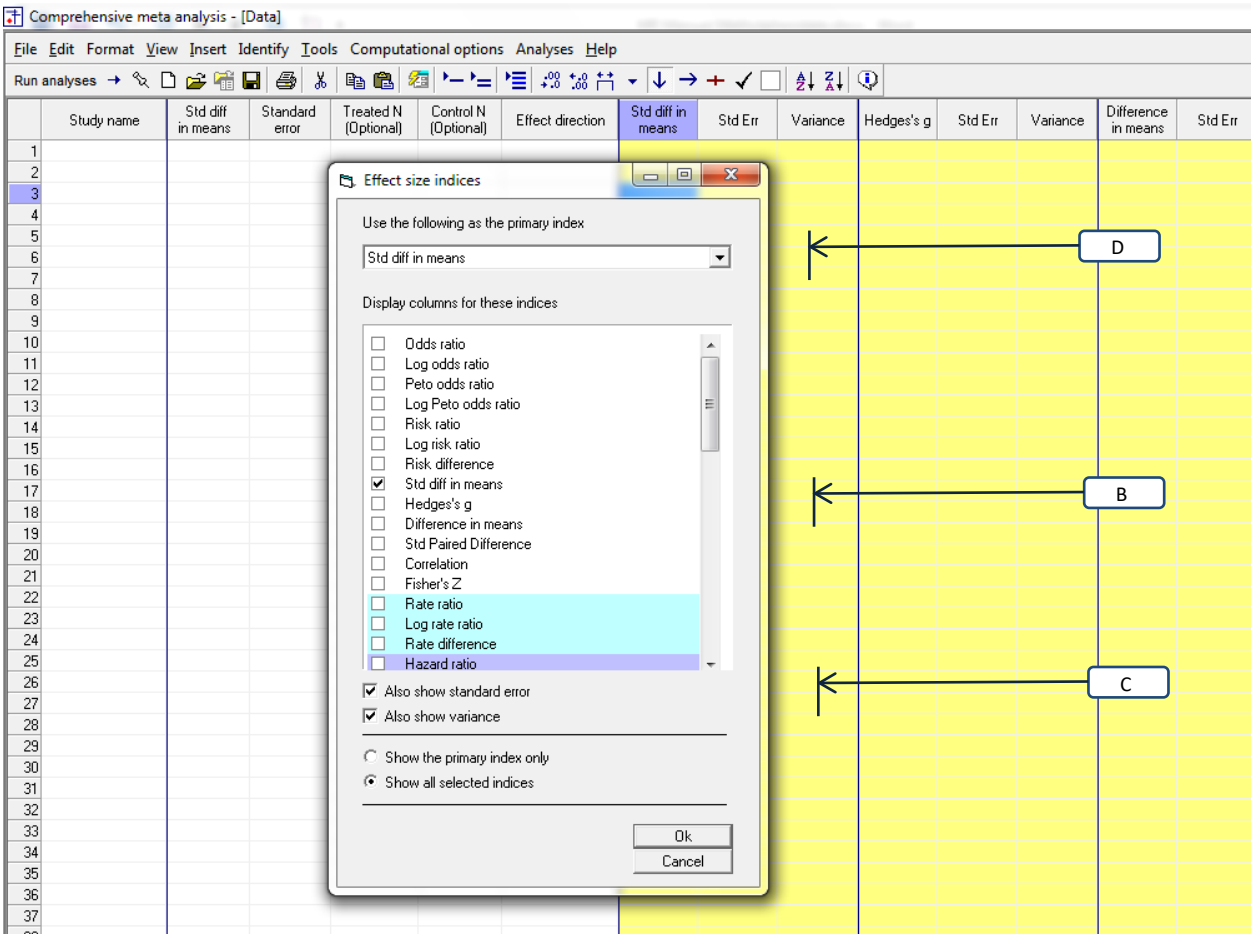


Figure 174 | Data-entry | Step 15

The screen now looks like Figure 175.

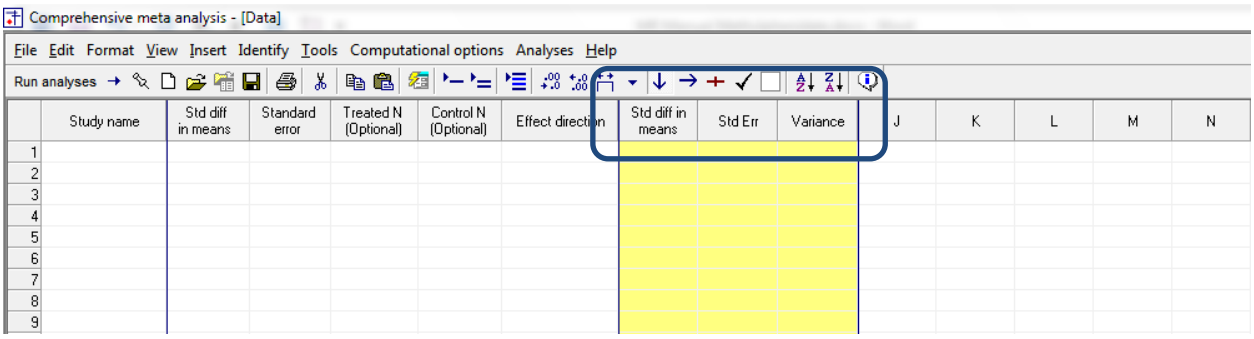


Figure 175 | Data-entry | Step 17

Insert columns for moderators (covariates)

Next, we need to create columns for the moderator variables. As shown in Figure 176,

- Click Insert > Column for > Moderator variable [L]

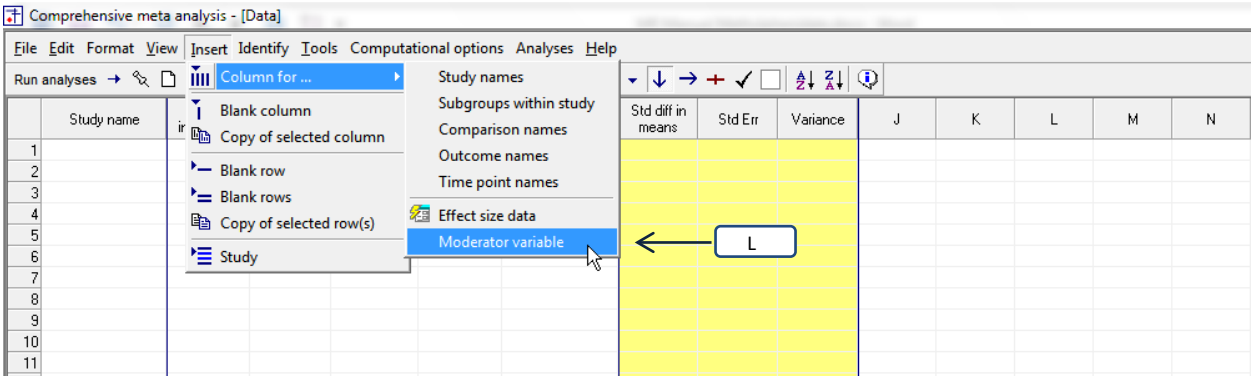


Figure 176 | Data-entry | Step 09

The program opens a wizard (Figure 177)

- Set the variable name to “Formulation” [M]
- Set the column function to Moderator [N]
- Set the data type to Categorical [O]

This will be coded “Continuous” for studies that employed a continuous formulation of the drug, and “Intermittent” for studies that employed a bi-phasic or other non-continuous formulation.

Then, click [Ok]

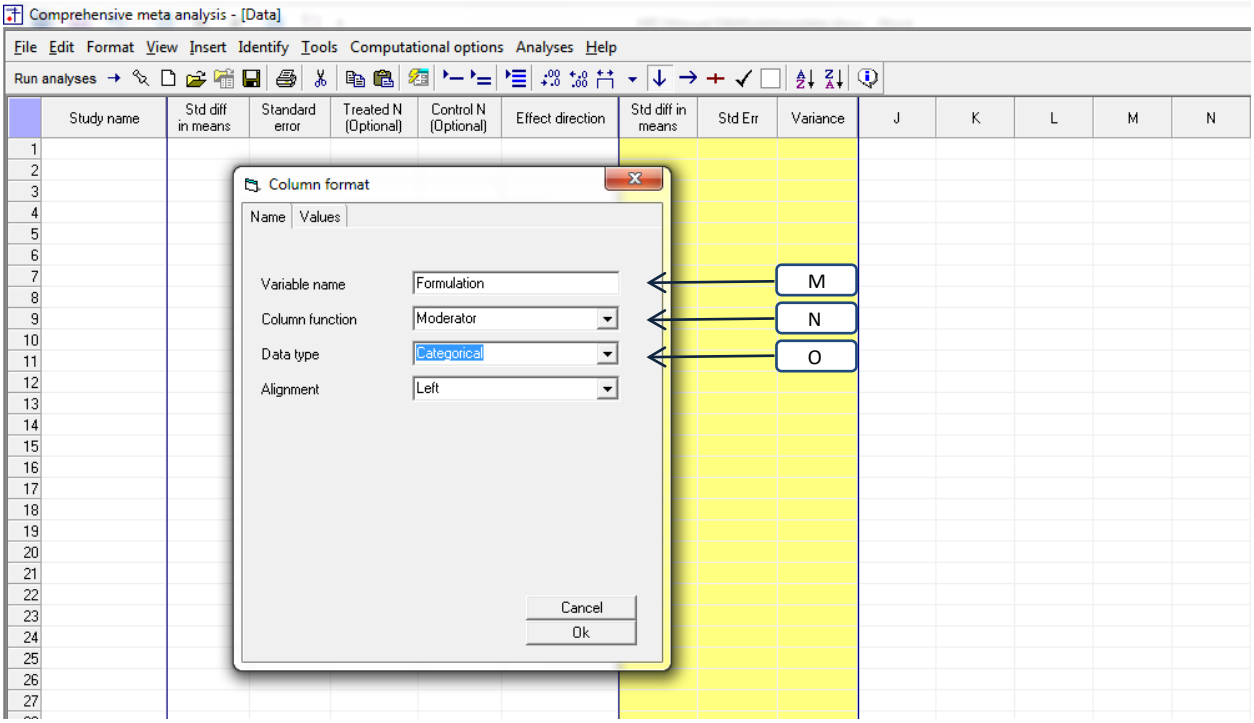


Figure 177 | Data-entry | Step 10

As shown in Figure 178, Click Insert > Column for > Moderator variable

- Set the variable name to “Dose”
- Set the column function to Moderator
- Set the data type to Decimal

Then, click [Ok]

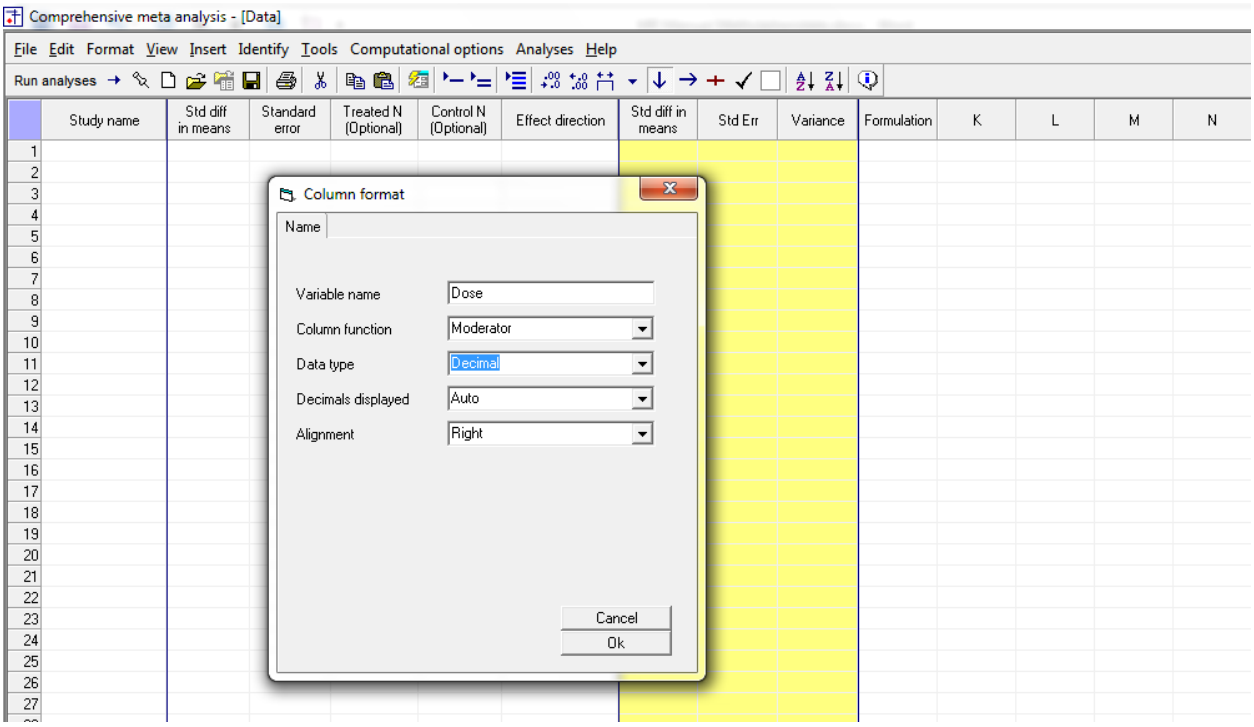


Figure 178 | Data-entry | Step 11

As shown in Figure 179, Click Insert > Column for > Moderator variable

- Set the variable name to “SUD”
- Set the column function to [Moderator]
- Set the data type to [Categorical]

This will be coded “Yes” for studies that enrolled patients with substance abuse disorder (SUD), and “No” for studies that excluded these patients.

Then, click [Ok]

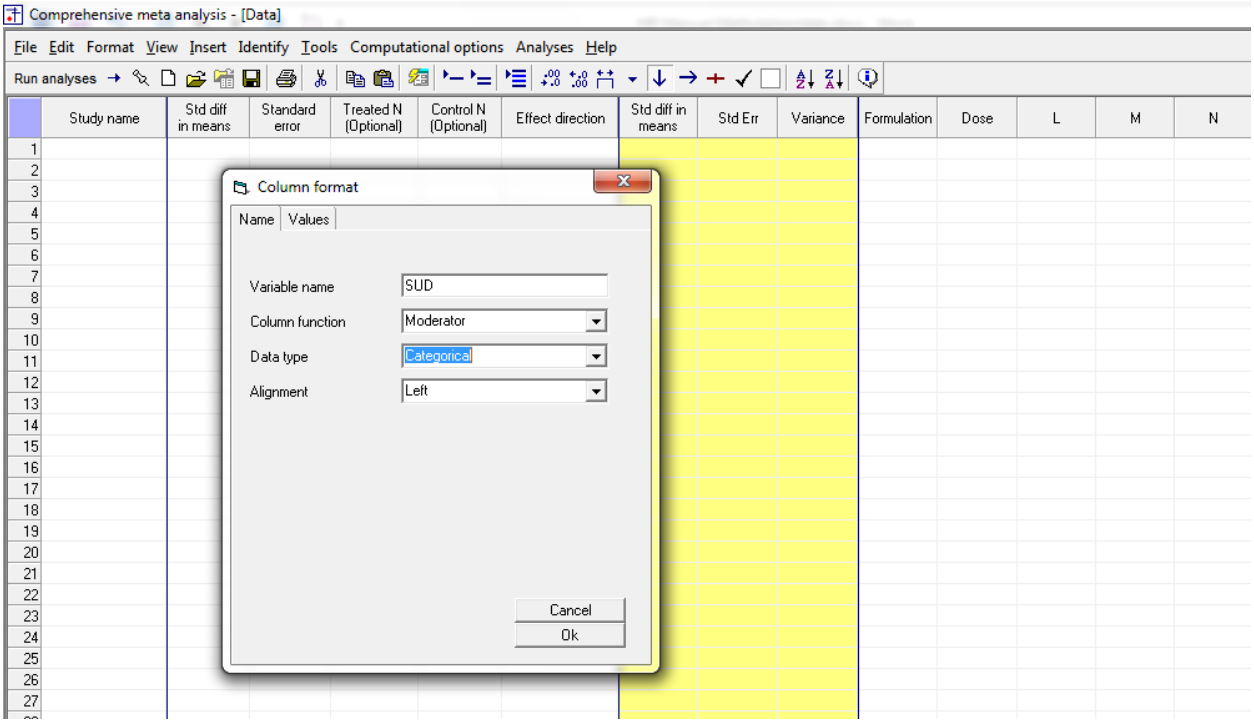


Figure 179 | Data-entry | Step 12

The screen is shown in Figure 180

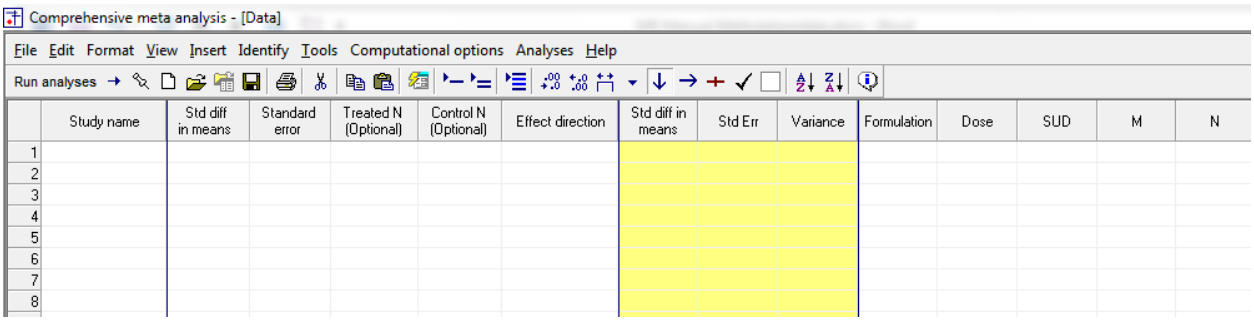


Figure 180

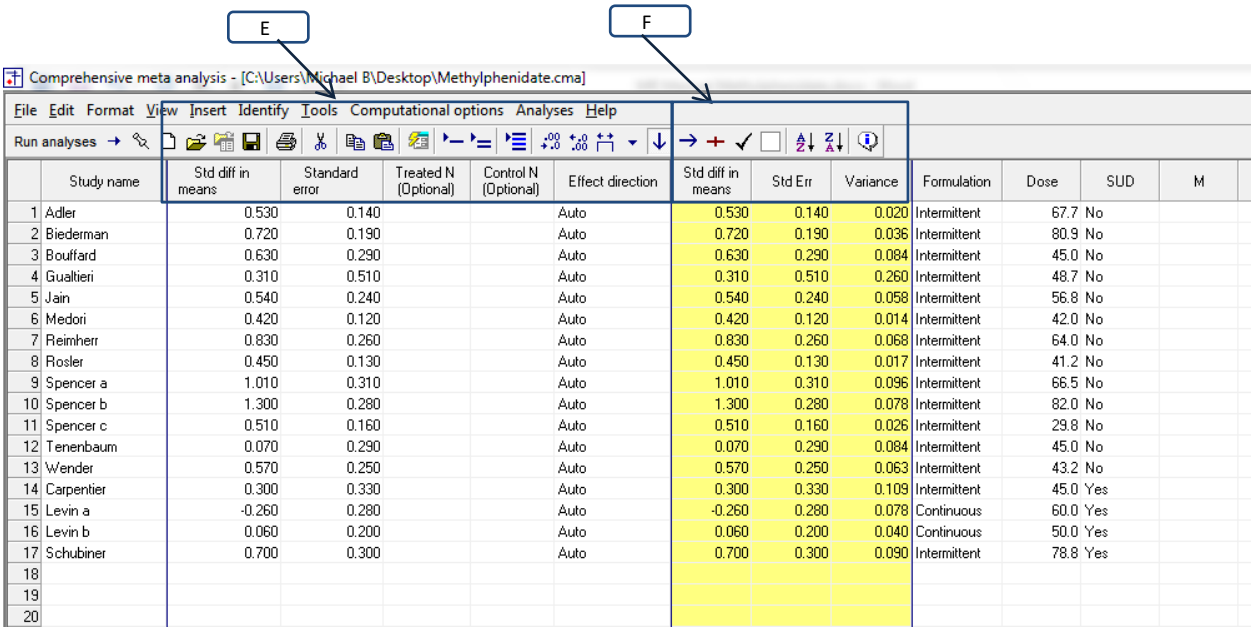
Enter the data

You can enter the data manually, or copy and paste from Excel ™ or another source (see Appendix 1: The dataset)

Figure 181 shows the data-entry sheet with the data for the 17 studies.

Typically, we enter summary data (such as means and standard deviations or events and N) into the white columns [E], and the program displays the computed effect sizes in the yellow columns [F]. In this example we entered the computed effect sizes into the white columns, and so the yellow columns simply duplicate this data.

Note that we've entered the standard difference in means and its standard error, since this is what was reported in the original paper. We don't have the sample size, so we leave that blank. The "Effect direction" is set to "Auto". This means that studies with an effect size above zero will be entered a positive numbers while studies with an effect size below zero will be entered as negative numbers. This refers to the effect size, and not to a test of significance.



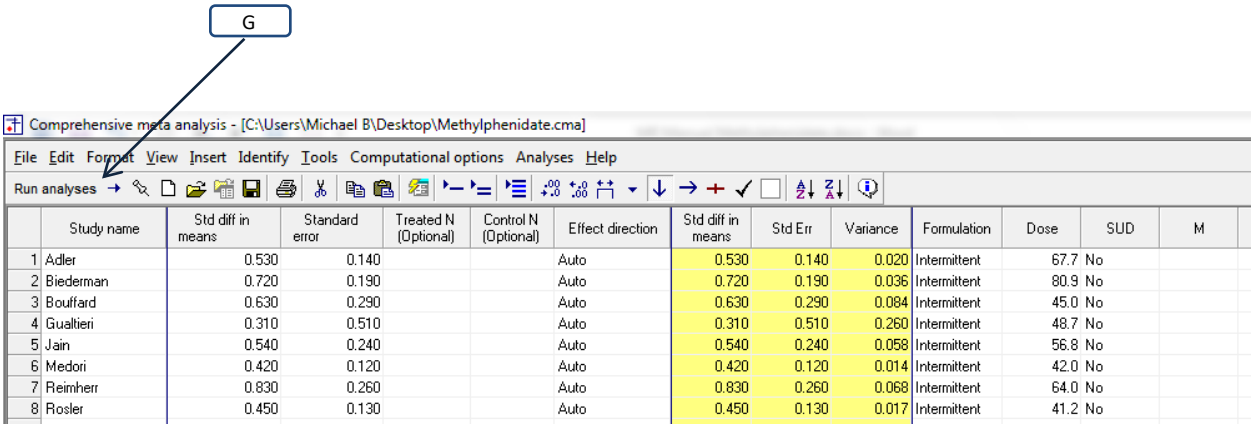
	Study name	Std diff in means	Standard error	Treated N (Optional)	Control N (Optional)	Effect direction	Std diff in means	Std Err	Variance	Formulation	Dose	SUD	M
1	Adler	0.530	0.140			Auto	0.530	0.140	0.020	Intermittent	67.7	No	
2	Biederman	0.720	0.190			Auto	0.720	0.190	0.036	Intermittent	80.9	No	
3	Bouffard	0.630	0.290			Auto	0.630	0.290	0.084	Intermittent	45.0	No	
4	Gualtieri	0.310	0.510			Auto	0.310	0.510	0.260	Intermittent	48.7	No	
5	Jain	0.540	0.240			Auto	0.540	0.240	0.058	Intermittent	56.8	No	
6	Medori	0.420	0.120			Auto	0.420	0.120	0.014	Intermittent	42.0	No	
7	Reimherr	0.830	0.260			Auto	0.830	0.260	0.068	Intermittent	64.0	No	
8	Rosler	0.450	0.130			Auto	0.450	0.130	0.017	Intermittent	41.2	No	
9	Spencer a	1.010	0.310			Auto	1.010	0.310	0.096	Intermittent	66.5	No	
10	Spencer b	1.300	0.280			Auto	1.300	0.280	0.078	Intermittent	82.0	No	
11	Spencer c	0.510	0.160			Auto	0.510	0.160	0.026	Intermittent	29.8	No	
12	Tenenbaum	0.070	0.290			Auto	0.070	0.290	0.084	Intermittent	45.0	No	
13	Wender	0.570	0.250			Auto	0.570	0.250	0.063	Intermittent	43.2	No	
14	Carpentier	0.300	0.330			Auto	0.300	0.330	0.109	Intermittent	45.0	Yes	
15	Levin a	-0.260	0.280			Auto	-0.260	0.280	0.078	Continuous	60.0	Yes	
16	Levin b	0.060	0.200			Auto	0.060	0.200	0.040	Continuous	50.0	Yes	
17	Schubiner	0.700	0.300			Auto	0.700	0.300	0.090	Intermittent	78.8	Yes	
18													
19													
20													

Figure 181 | Data-entry | Step 18

You may enter (or copy and paste) the data as explained in the appendix. Or, simply open the file ADHD.cma. There are two versions of this file, one using a period to indicate decimal places and one using a comma. Use the one that corresponds to your computer's settings.

STEP 2: RUN THE BASIC META-ANALYSIS

To run the analysis, click [Run Analysis] as shown in Figure 182 [G].



	Study name	Std diff in means	Standard error	Treated N (Optional)	Control N (Optional)	Effect direction	Std diff in means	Std Err	Variance	Formulation	Dose	SUD	M
1	Adler	0.530	0.140			Auto	0.530	0.140	0.020	Intermittent	67.7	No	
2	Biederman	0.720	0.190			Auto	0.720	0.190	0.036	Intermittent	80.9	No	
3	Bouffard	0.630	0.290			Auto	0.630	0.290	0.084	Intermittent	45.0	No	
4	Gualtieri	0.310	0.510			Auto	0.310	0.510	0.260	Intermittent	48.7	No	
5	Jain	0.540	0.240			Auto	0.540	0.240	0.058	Intermittent	56.8	No	
6	Medori	0.420	0.120			Auto	0.420	0.120	0.014	Intermittent	42.0	No	
7	Reimherr	0.830	0.260			Auto	0.830	0.260	0.068	Intermittent	64.0	No	
8	Rosler	0.450	0.130			Auto	0.450	0.130	0.017	Intermittent	41.2	No	

Figure 182 | Data-entry | Step 19

The main analysis screen

The program displays the main analysis screen (Figure 183).

The current effect size [A] is “Std diff in means”.

The next few pages outline the main analysis in *CMA* using the traditional approach. However, this is optional. As soon as you arrive at the main analysis screen (Figure 183) you can click [Analysis > Meta-regression 2] to proceed immediately to the regression module.

The initial meta-analysis

In Figure 183, the <Fixed> tab [B] is selected, so the program is displaying the results for a fixed-effect analysis [C].

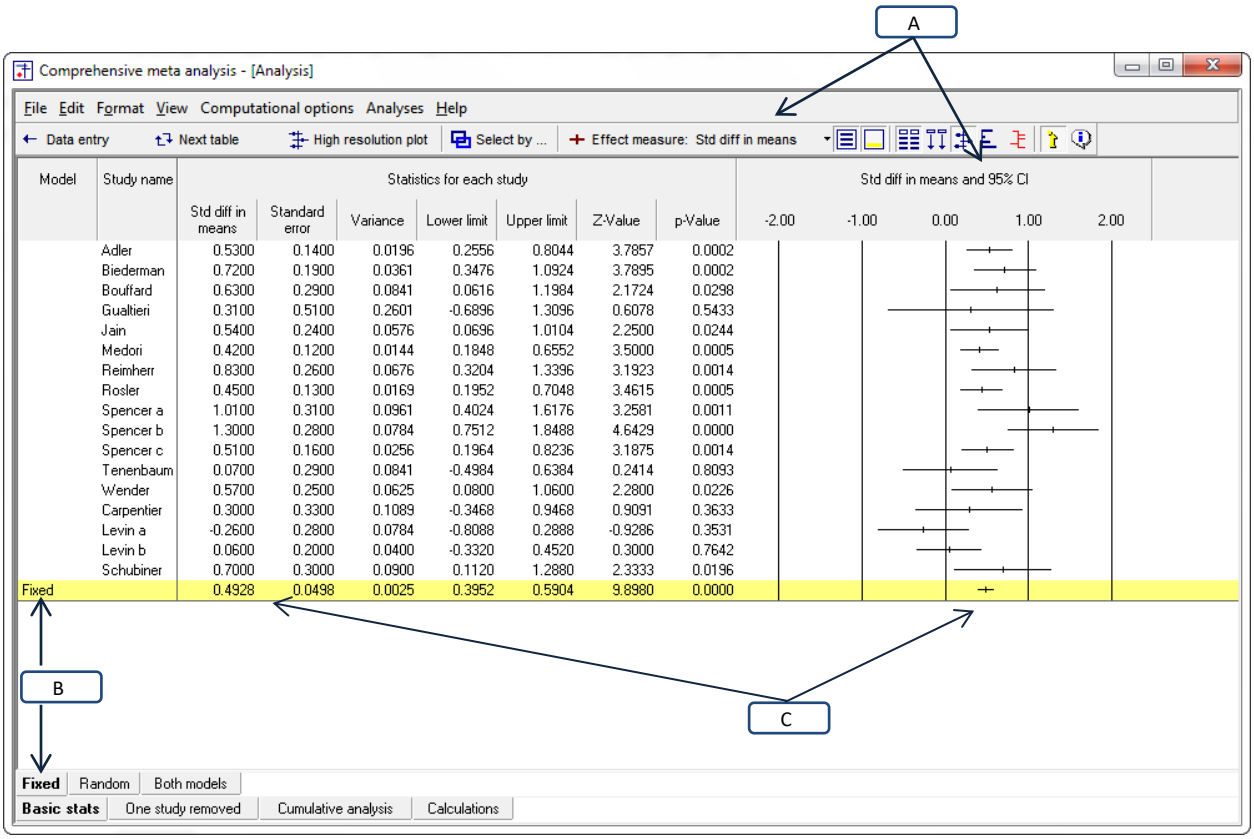


Figure 183 | Basic analysis | Fixed-effect

Click the tab [D] for <Random>. The program [E] displays results for a random-effects analysis.

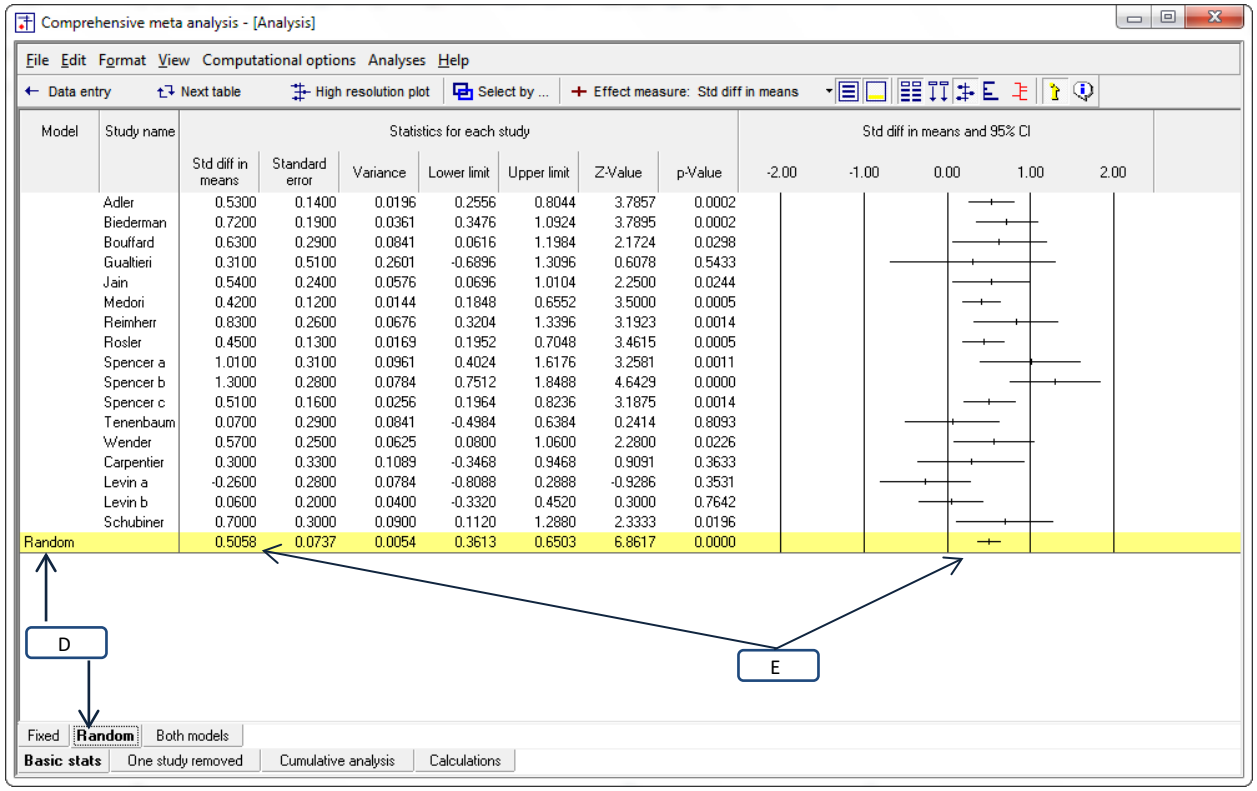


Figure 184 | Basic analysis | Random-effects

Display statistics

Click <Next table> to display the statistics shown in Figure 185.

Using random-effects weights [J], the mean effect size (d) is 0.5058. The *Z*-value is 6.8617 with a corresponding *p*-value of < 0.0001. Thus, we can reject the null hypothesis that mean effect size is zero, and conclude that the drug is more effective than placebo.

At the same time, there is also a substantial amount of dispersion in the effect size. As shown in the figure [K], the variance of the true effects (T2) 0.0309 and the standard deviation of true effects (T) is 0.1966. To get a general sense of the true dispersion we can assume that the true effects are balanced about the random-effects estimate of the mean effect, and that some 95% of all true effects fall within approximately 2*T* of this mean. Then most true effects fall in the range of 0.11 to 0.73.

From a clinical perspective this is a wide range and it would be important to understand the reason for this dispersion.

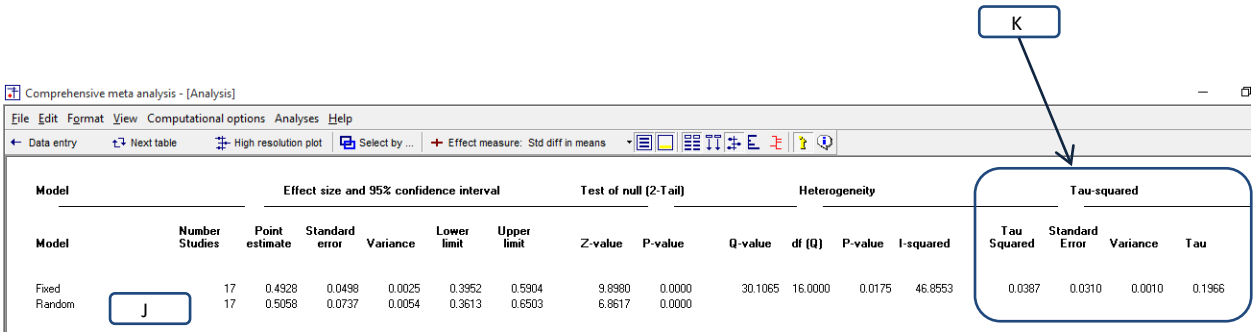


Figure 185 | Basic analysis | Display statistics for heterogeneity

Display moderator variables

It's always a good idea to get a visual sense of the data before proceeding to the analyses. For this purpose we'll plot the data as a function of the putative risk factors, which are dose, SUD, and formulation. Note that this is optional, and has no effect on the regression.

In Figure 186, click View > Columns > Moderators [F]

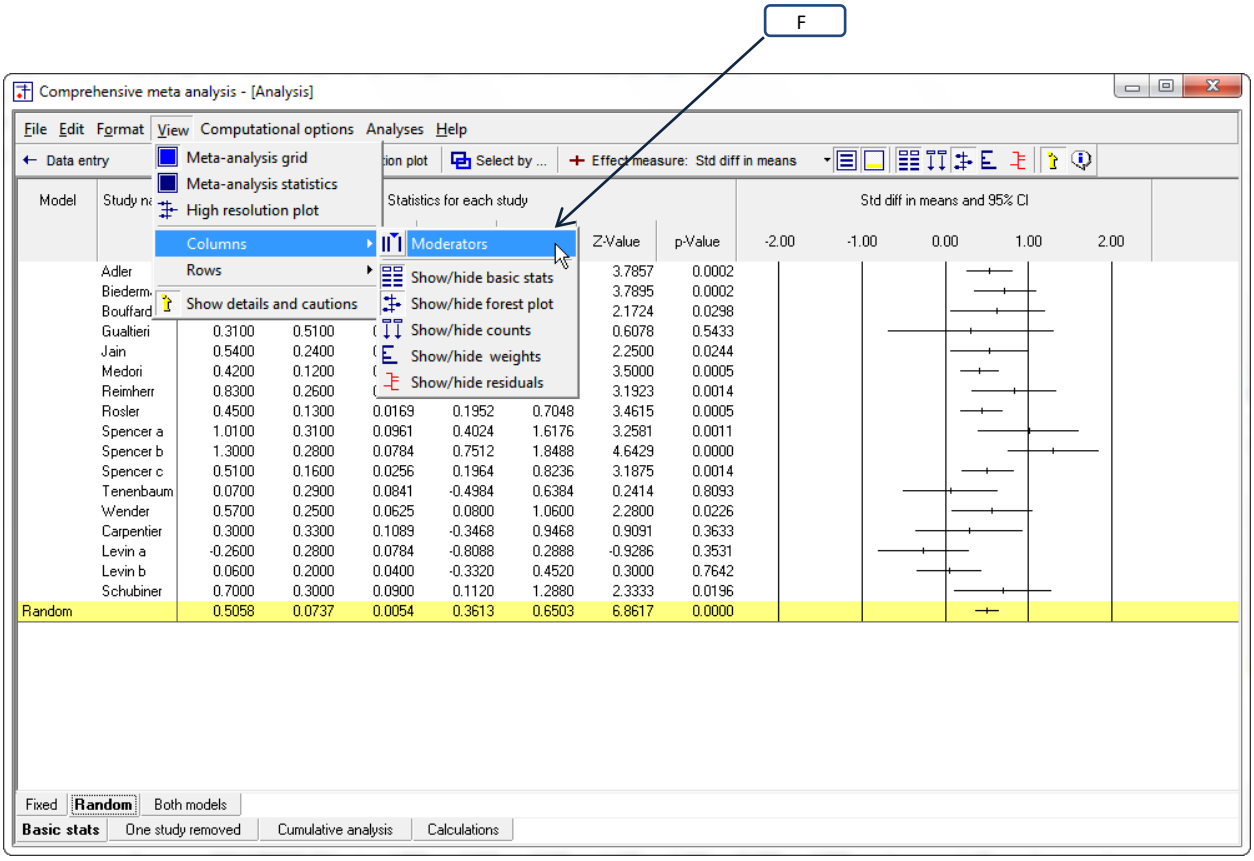


Figure 186 | Basic analysis | Display moderators

The program displays a list of all variables that had been defined as moderators on the data-entry screen.

In Figure 187 [G], drag and drop Dose onto the main screen, to the right of the "*p*-value" column.

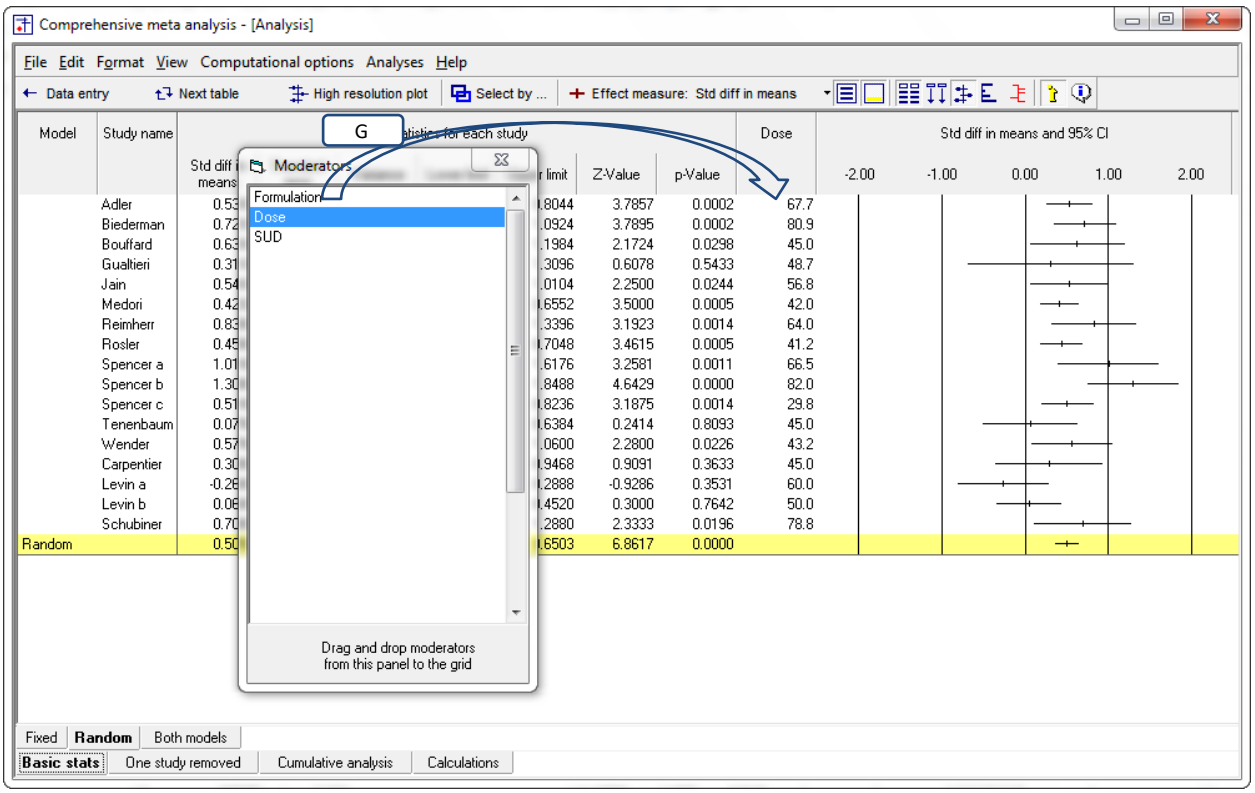


Figure 187 | Basic analysis | Display moderators

The screen should now resemble Figure 188, with Dose displayed [H].

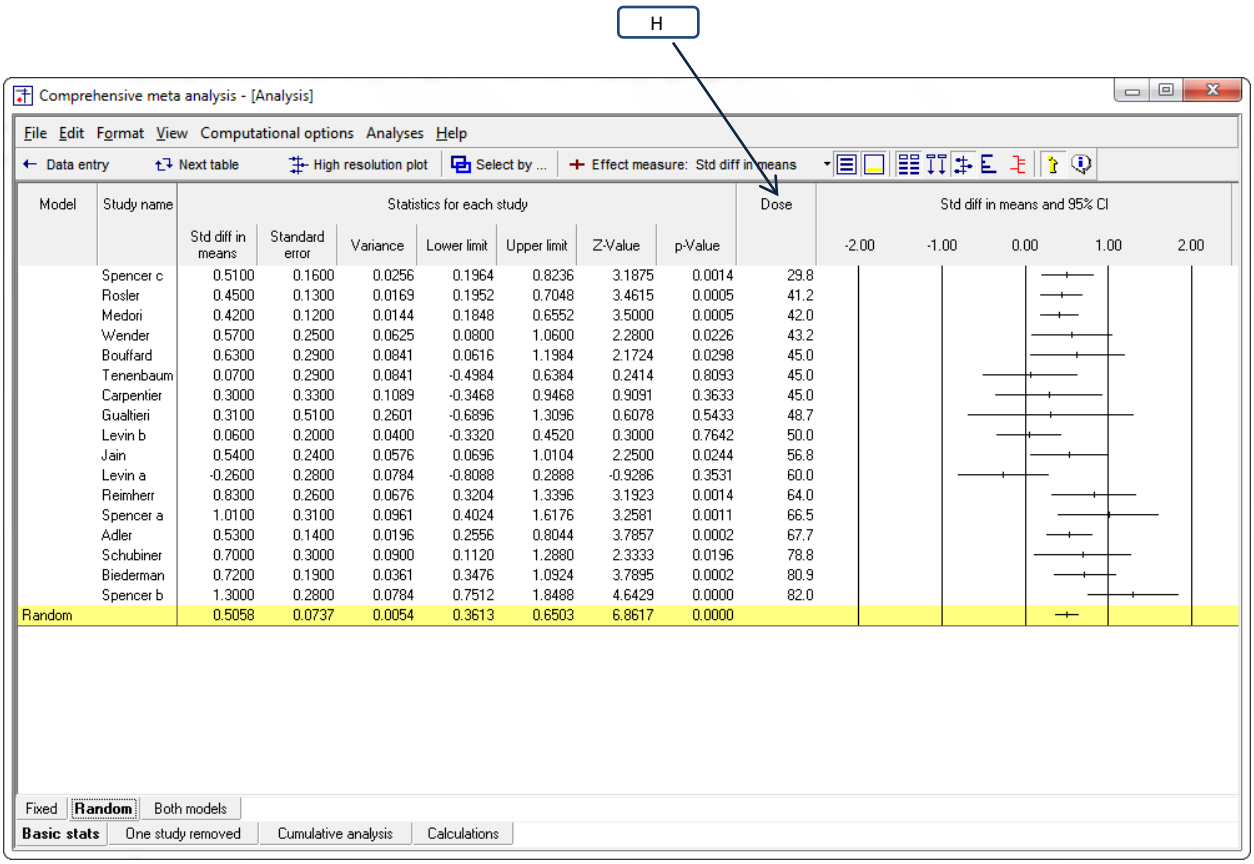


Figure 188

You can right-click on any column and sort by that column. Here, the studies are sorted by dose [I].

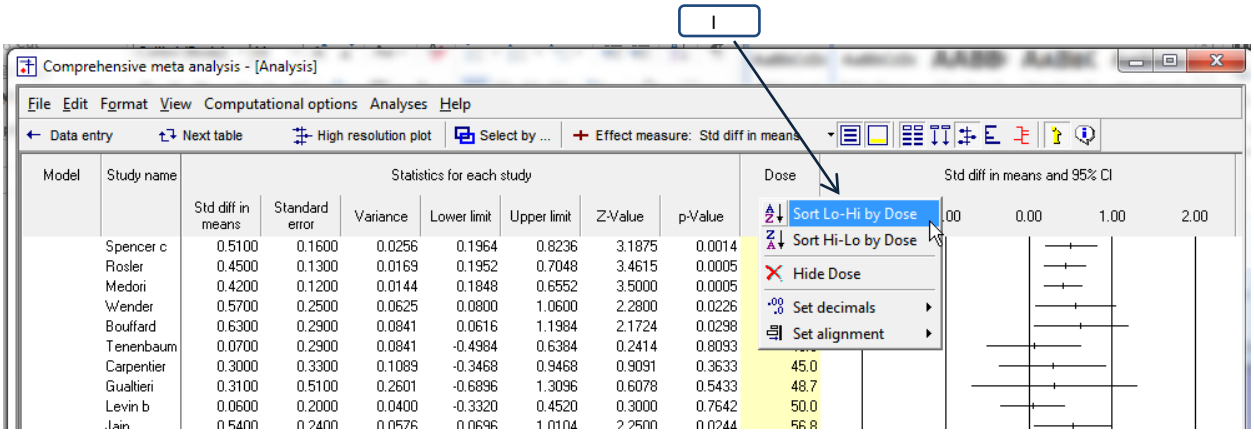


Figure 189

It appears that the effect size tends to be modest (near 0.50) for in the top third of the plot (doses in the range of 30 to 45) [J], smaller for those in the middle-third of the plot (doses in the range of 45 to 60) [K] and largest for those in the bottom third of the plot (doses in the range of as high as 1.30) [L].

We don't know yet if this pattern is real (or if it can be explained by sampling error). If it does turn out to be real, we need to consider that it could be due to confounds with other risk factors. We would also tend to be more skeptical if we had not anticipated this pattern in advance. With these important caveats in mind, we'll show how we can use regression to try and sort out these issues.

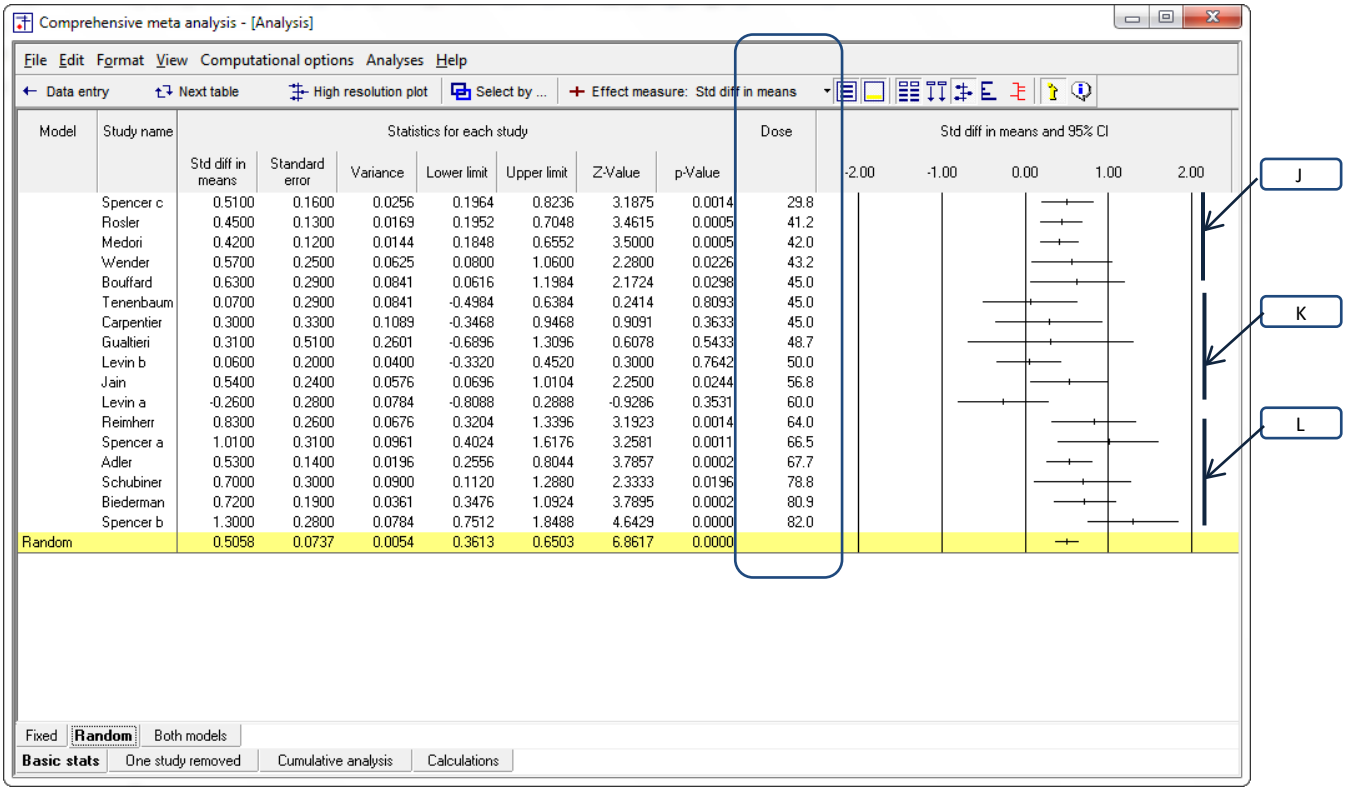


Figure 190 | Basic analysis | Display moderators

Add covariates to the model

When you initially open the regression module the program displays the following

- The main screen [A]
- A list of available covariates [B]

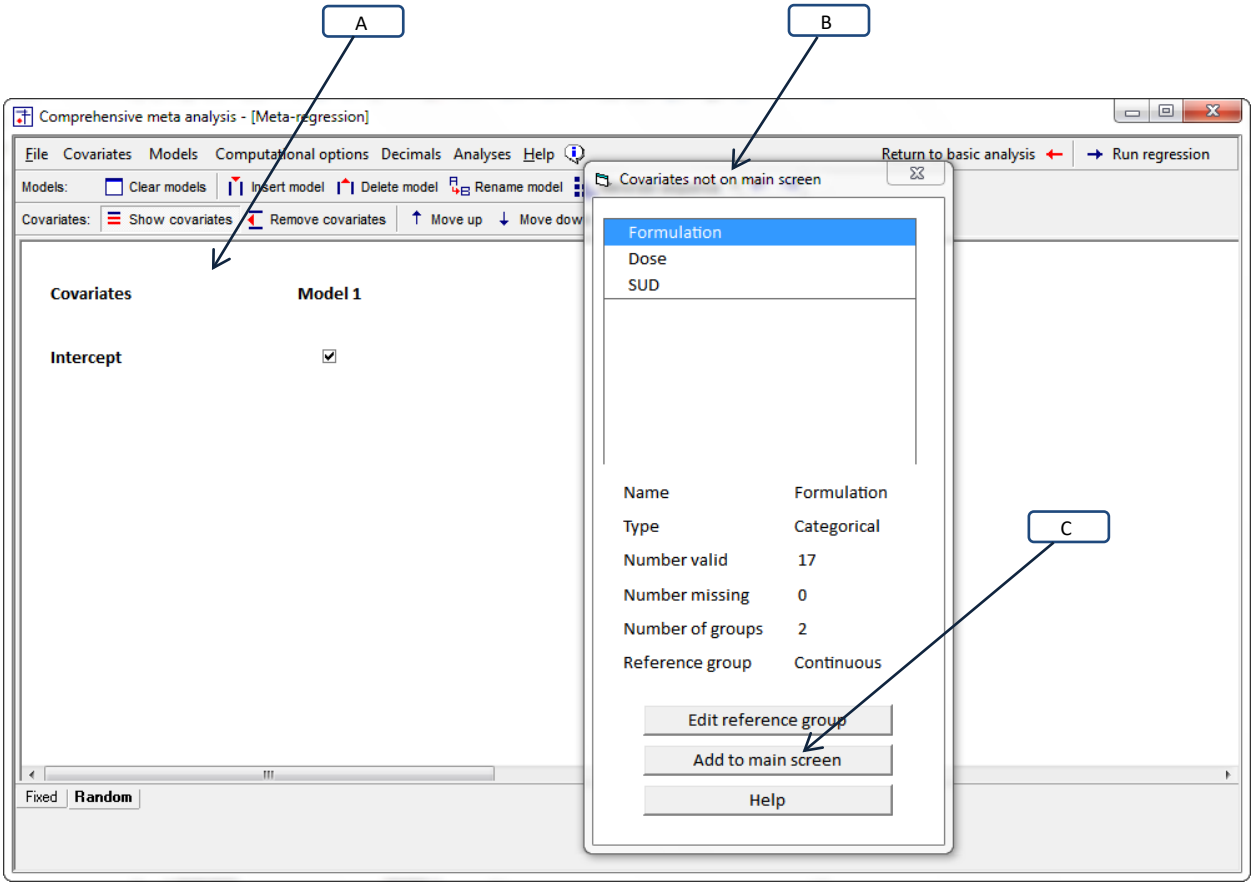


Figure 191 | Run regression | Step 03

To add a continuous covariate

To add a categorical covariate

Chapter ___ provides a discussion of how to select the reference group

We want to move Dose from the wizard [B] onto the main screen [A].

- Click "Dose" on the wizard CHANGE PICTURE
- Click [Add to main screen] [C]
- Tick the box for Dose
- SHOW

-
-

The model is shown in In Figure 192 [D].

Tick the check-boxes for all covariates [E]

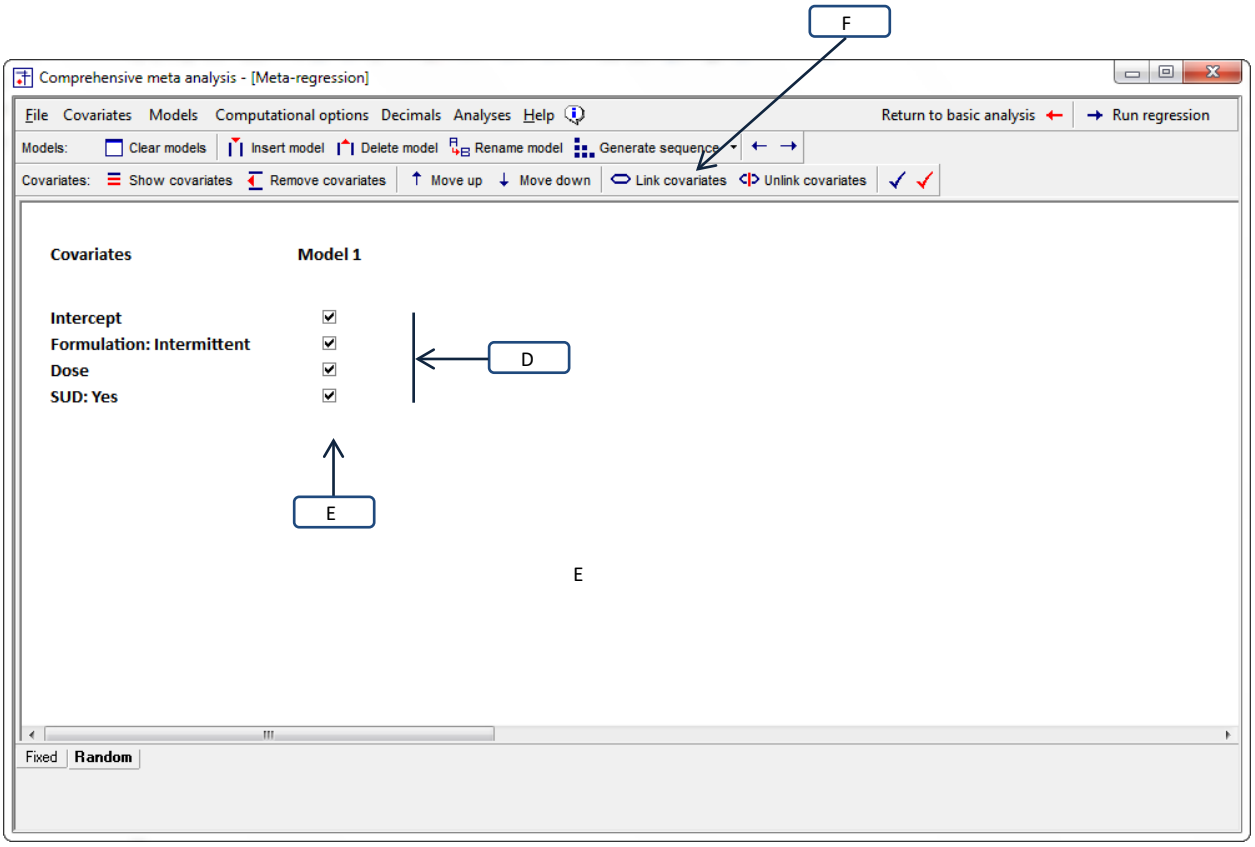


Figure 192 | Run regression | Step 04

The covariates are controlled by the “Covariates” toolbar [F]. On this toolbar,

- [Show covariates] shows or hides the wizard
- [Remove covariates] allows you to remove a covariate from the main screen
- [Move up] and [Move down] allow you to edit the sequence of covariates
- The blue and red checks allow you to add (or remove) checks from a series of check-boxes

To add categorical covar

Set computational options

The program allows you to specify various options for the computations

Click Computational options to display the menu in Figure 193.

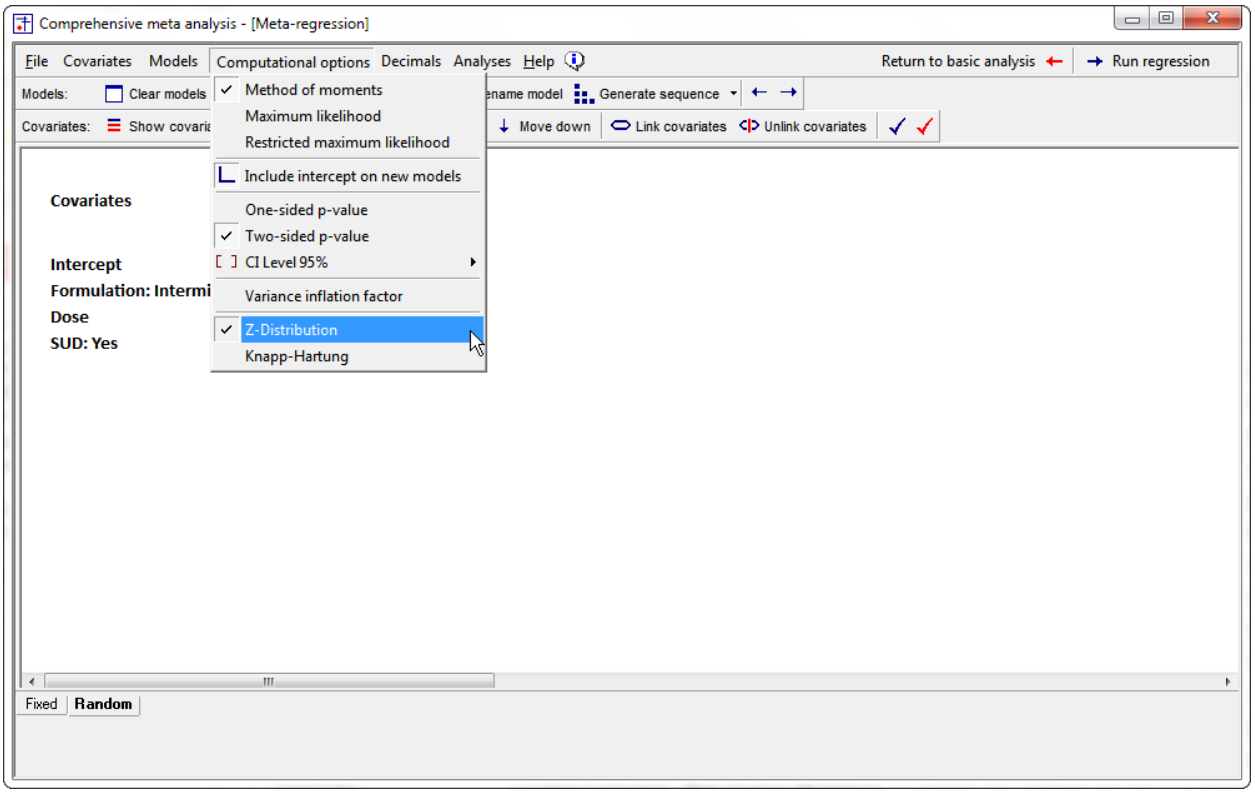


Figure 193 | Run regression | Step 05

Each of these options is discussed in Computational options.

Set the options as follows.

- Method for estimating T^2 > Method of moments
- Include intercept on new models > Checked
- One-tailed or two-tailed test for *p*-values > Two-tailed
- CI level > 95%
- Variance inflation factor > Unchecked
- *Z* distribution or Knapp-Hartung > *Z* distribution

Run the regression

To run the regression, simply click “Run regression” on the toolbar [A] in Figure 193

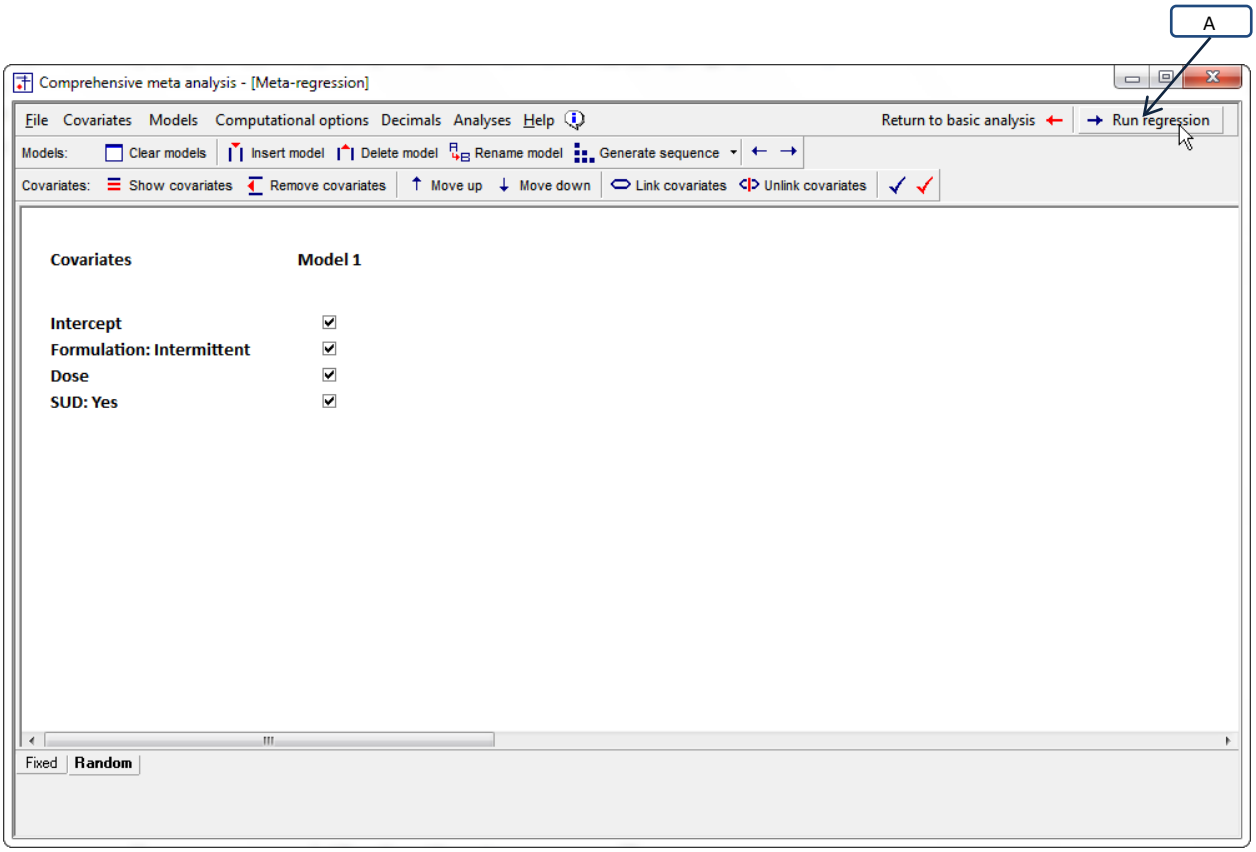


Figure 194 | Run regression | Step 06

Other screens

MOVE

To navigate to other tables of results, click “More results” [A] in Figure 195 and then select any of the following.

About the predictive model

- Covariance (see page 103)
- Correlation (see page 104)
- Diagnostics (see page 74)
- *R*-squared graphic (see page 68)

About the data included in (or excluded from) the analysis

- All studies (see page 233)
- Valid studies (see page 233)

Statistics that compare different predictive models

- Increments (see page **Error! Bookmark not defined.**)
- Models summary (see page 239)
- Compare models (detailed) (see page 239)
- Compare models (*p*-values) (see page 239)

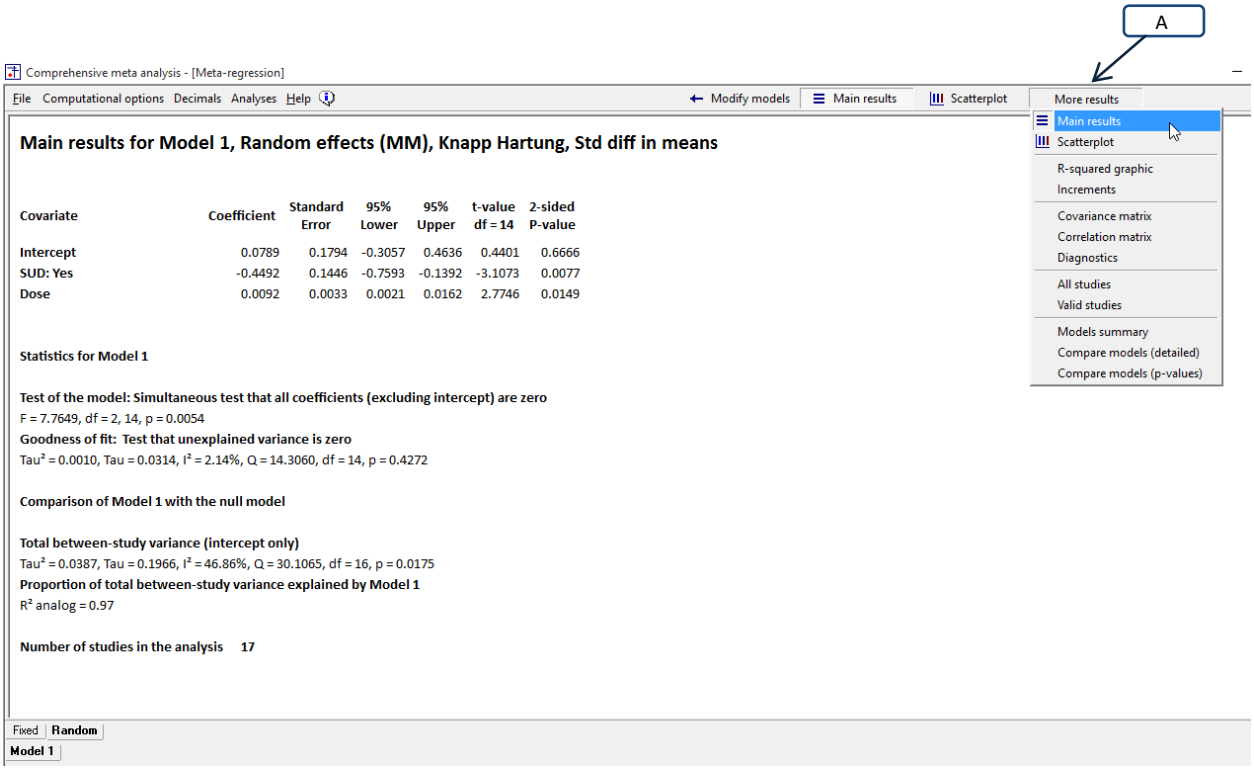


Figure 195 | Other screens

WORKING WITH THE PLOT

To create a plot, run a regression and then click [Scatterplot] [A]

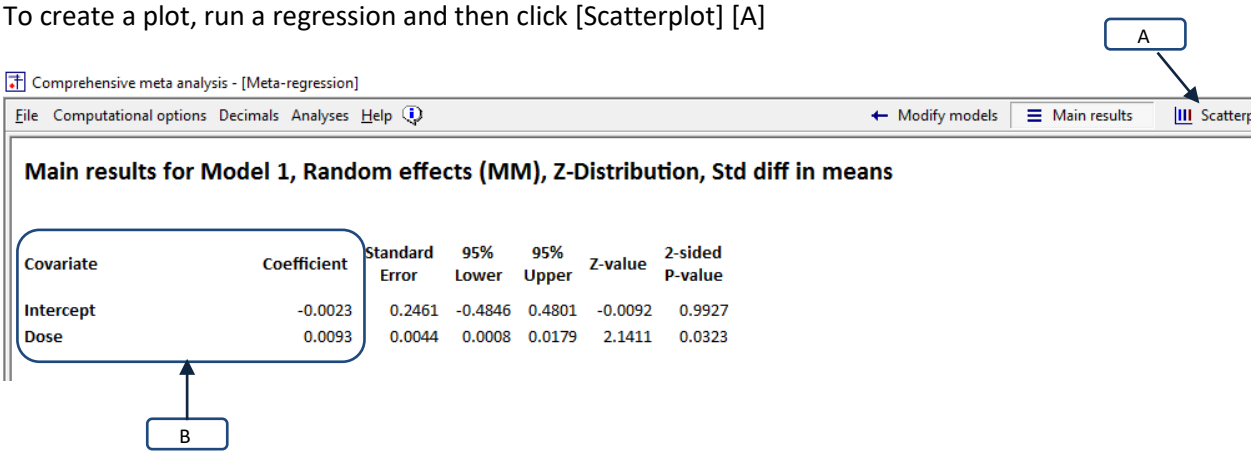
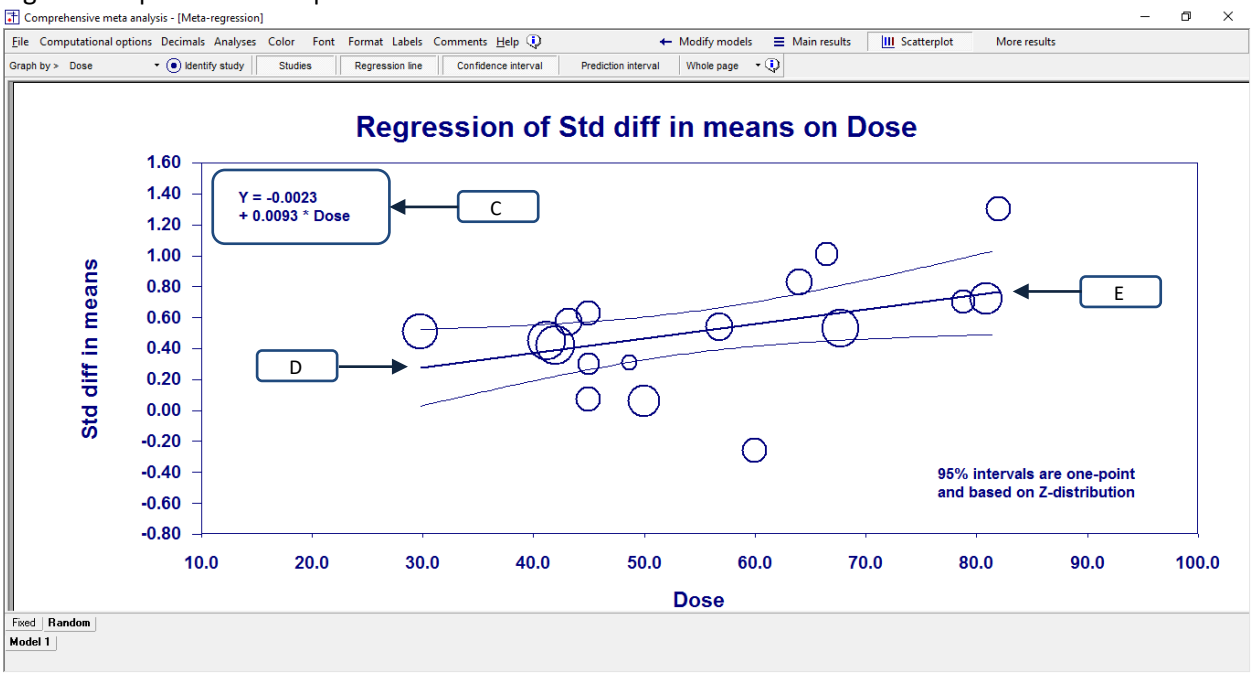


Figure 196 | Main results | Random-effects



The prediction equation is shown in both the main results [B] and the plot [C] as

$$Y = -0.0023 + 0.0093 \times Dose\,, \tag{1.122}$$

which tells us that as Dose increases by one unit, the predicted effect size (*d*) will increase by 0.0093. For a dose of 30 [D] the predicted effect size is

$$Y = -0.0023 + 0.0093 \times 30 = 0.1865\,, \tag{1.123}$$

and for a dose of 80 [E] the predicted effect size is

$$Y = -0.0023 + 0.0093 \times 80 = 0.8756 \,. \tag{1.124}$$

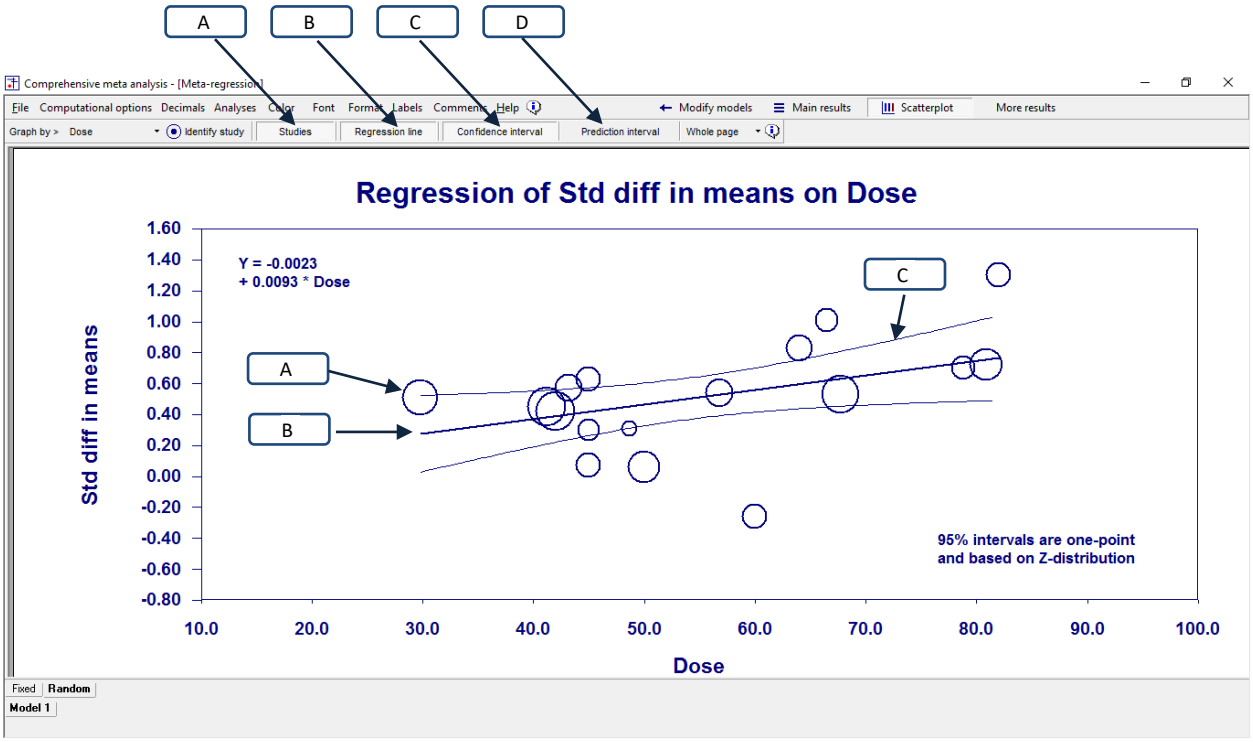
The regression line in the plot connects these two points. Note that the difference between 0.1865 and 0.8756 is 0.6900, which is simply 50 times 0.0093.

The 95% confidence interval for the coefficient is 0.0008 to 0.0179. The true regression line could be substantially more shallow or more steep than the one we've plotted, but it is probably not zero.

Main elements

The main elements on the plot are

- The studies [A]
- The regression line [B]
- The confidence interval [C]
- The prediction interval [D]

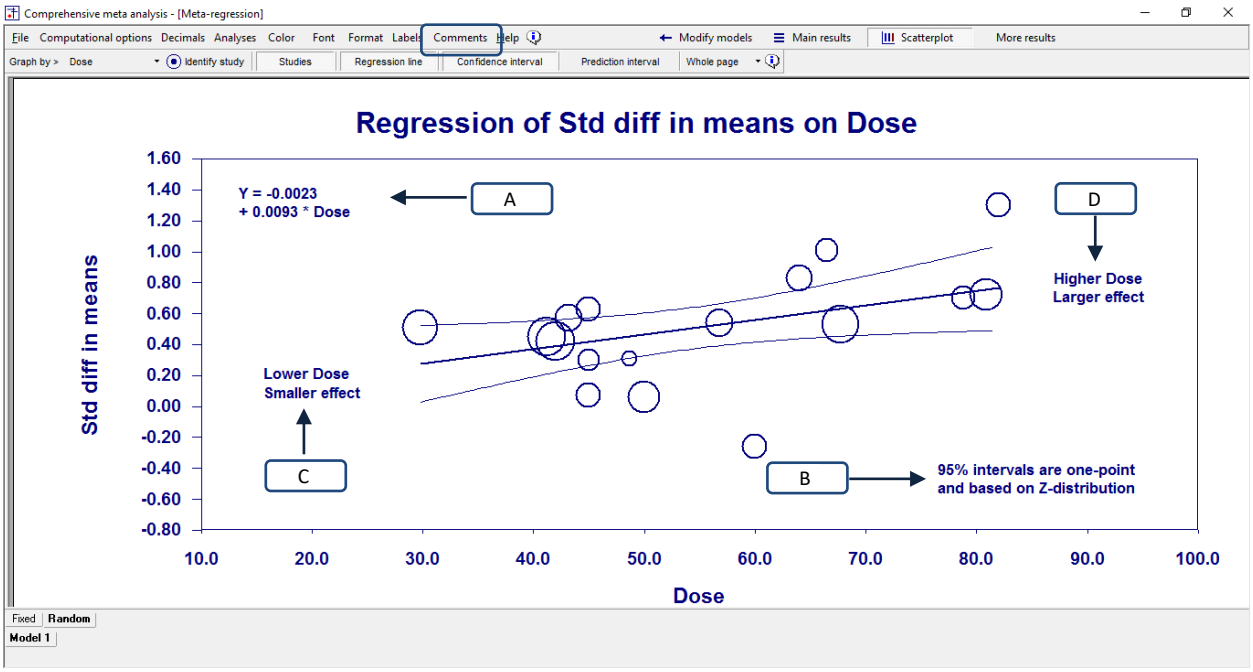


- To show or hide any of these elements, click the corresponding tools on the toolbar.
- To change the appearance of the studies [A], click Format > Studies.
- To change the appearance of the lines [B], [C], or [D], click Format > Lines

Annotations

You can add any of the following annotations

- The prediction equation [A]
- The statistical options [B]
- Comment-1 [C]
- Comment-2 [D]



The regression equation [A] shows the equation being used to draw the regression line. You can edit the number of decimals, the format of the equation, and its position on the page.

The statistical options notation [B] shows the statistical options in effect. To modify these options, use the Computational options menu.

The optional comments [C] and [D] can be used for any purpose.

- To show or hide any of these elements, click Comments > Item > Show/Hide
- To edit any item, click Comments > Item > Edit.
- To position any of these on the screen, click Comments > Item > Edit. First select the basic position (Upper left, Upper right, and so on). Then use the tools to move the item horizontally or vertically.

To identify specific studies

The program allows you to identify any study in the plot as shown in _____

- Click [Identify study] [A]
- Click on any study [B]
- The program displays the study name [C]

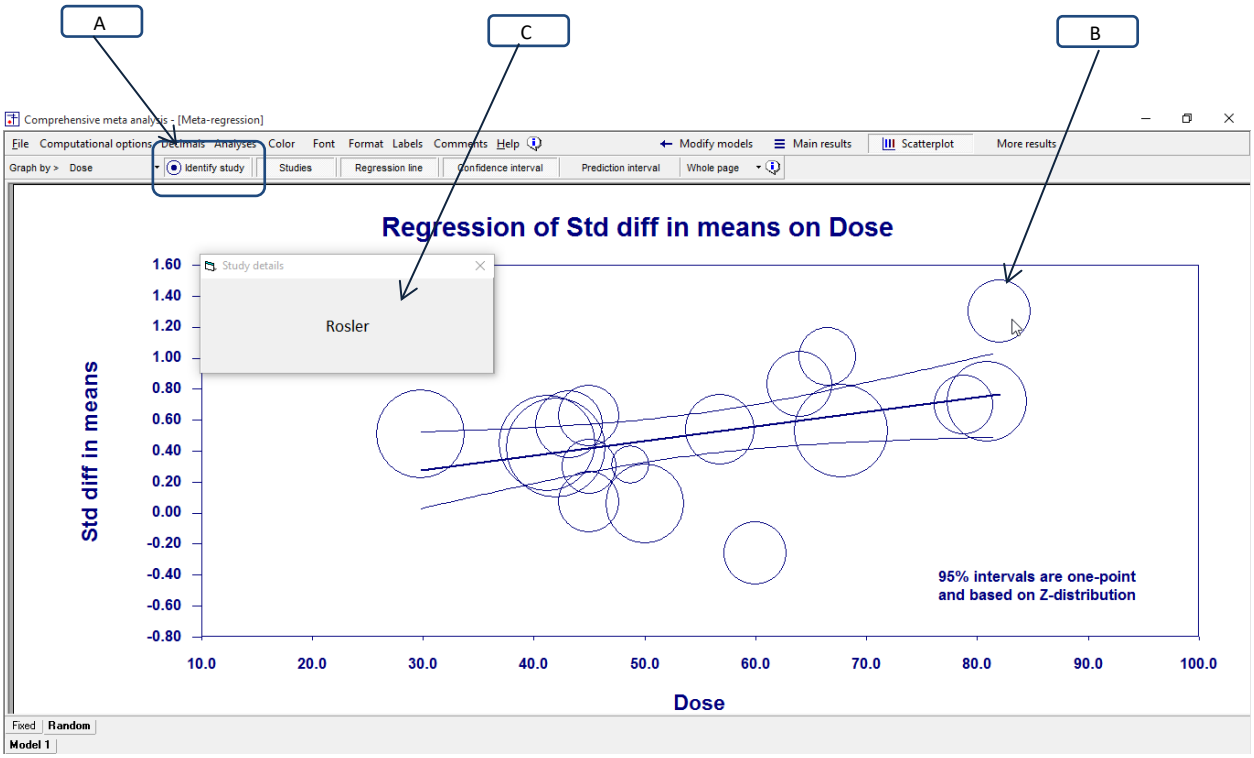


Figure 197 | Plot of log risk ratio on Dose | Identify studies

CONFIDENCE INTERVAL AND PREDICTION INTERVAL

Click [Confidence interval] to display/hide the confidence interval [A].
Click [Prediction interval] to display/hide the prediction interval [B].

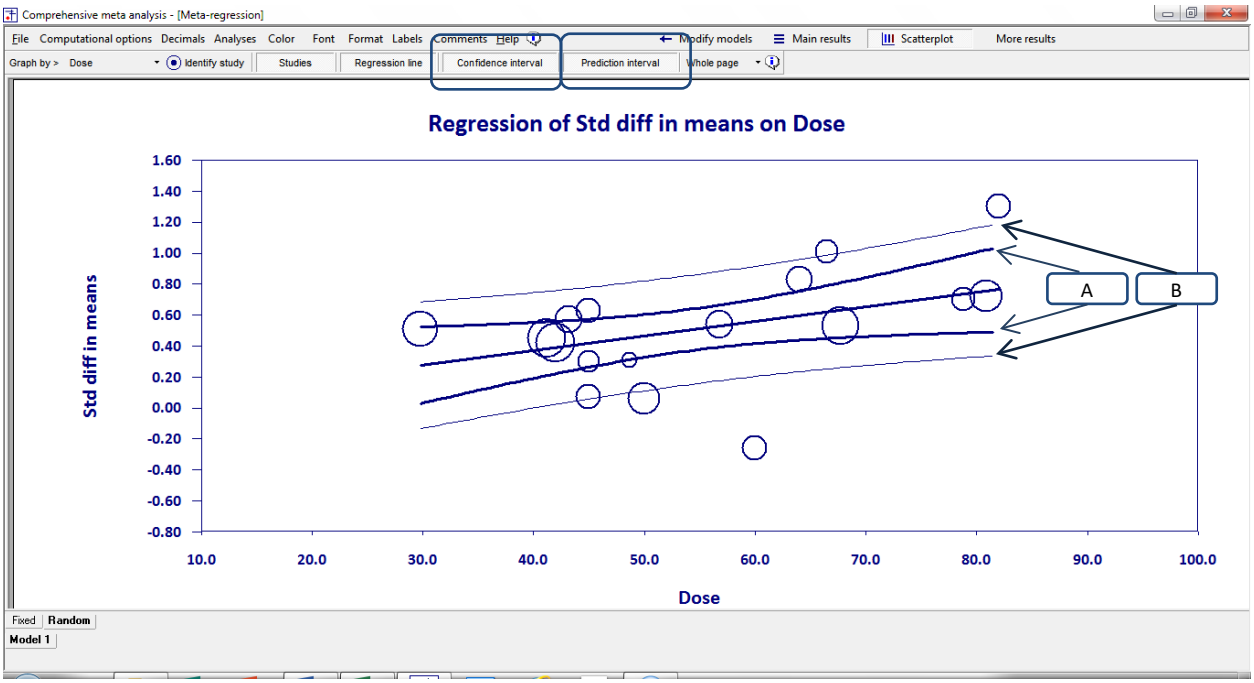


Figure 198 | Plot of effect size on Dose | Prediction interval

The confidence interval

The confidence interval is an index of precision that tells us how precisely we are able to estimate the mean effect. It is based (approximately) on the coefficient plus or minus two standard error. As such, the interval depends strongly on the number of studies in the analysis. With more studies we are able to estimate the coefficient more accurately.

The prediction interval is an index of dispersion that tells us how widely the effects are distributed. It is based (approximately) on the coefficient plus or minus two standard deviations. As such, (for the most part) the interval does not depend on the number of studies in the analysis. If the actual effects are distributed over a given range, the number of studies we choose to include in our analysis has no impact on that range.

For example, consider the effect size for studies with a Dose of 80 units. Numbers below are read from the graph and are approximate.

- The predicted effect size is 0.75. This is the *estimated mean* for all possible studies with this Dose.

- The confidence interval [A] is 0.45 to 1.05. The *actual mean* for all possible studies with this Dose probably falls in this range.

- The prediction interval [B] is 0.30 to 1.20. If we randomly select 100 studies with this Dose, 95 of them will probably have a true effect size in this range.

When working with the confidence interval or the prediction interval we need to base the intervals on one-point or simultaneous computations.

- One-point – In 95% of analyses, the confidence interval *at any single dose* will include the true mean effect for that dose.
- Simultaneous – In 95% of analyses, the confidence interval *at all doses* will include the true mean effect for those doses.

- One-point – In 95% of analyses, the prediction interval *at any single dose* will include the true effect for a study selected at random at that dose.
- Simultaneous – In 95% of analyses, the prediction interval *at all doses* will include the true effect for a study selected at random at that dose.

Use the computational options menu to select between these two options, and also to set the confidence level.

Other options for customizing the graph are as follows

Appearance		
Line width	Format > Line width	
Font	Font	
Font size	Format > Font size	
Title and labels		
Title	Labels > Title	
X-Axis	Labels > X-axis	
Y-Axis	Labels > Y-axis	
Study circles		
Proportionate	Format > Studies	
Line width	Format > Studies	
Axes		
Scale for X-axis	Format > X-axis	
Scale for Y-axis	Format > Y-axis	
Decimals	Format > Decimals	
Statistical Model		
Fixed	Select Fixed tab at bottom of screen	
Random	Select Random tab at bottom of screen	
Predictive model		
Model 1	Select desired model at bottom of screen	
Export		
To Word	Files > Export to Word	
To PowerPoint	Files > Export to PowerPoint	
To File	Files > Export to File	
To Clipboard	Files > Copy to clipboard	
Comments		
Equation	Show / Hide / Edit	(The prediction equation)
Annotation	Show / Hide / Edit	(For the confidence interval and prediction interval)
Comment 1	Show / Hide / Edit	(User's optional comment)
Comment 2	Show / Hide / Edit	(User's optional comment)

Decimals	
Equation in plot	Select number
X-Axis	Select number
Y-Axis	Select number

- To set the circles to be proportionate to the study weight (or not) Click Format > Studies
- To edit the appearance of the circles Click Format > Studies
- To modify the color of the circles Click Color > Edit colors > Studies
-
- To edit the appearance of the regression line Click Format > Regression line
- To modify the color of the regression line Click Color > Edit colors > Regression line

Modify the colors

The program maintains two color schemes. These are called Printing and PowerPoint but can actually be used for any purpose. To switch between the schemes click

- Color > Use colors for printing
- Color > Use PowerPoint

After you've selected one scheme or the other, you can edit the color for any element on the screen. To modify colors for the current color scheme click

- Color > Edit colors (for current scheme)

Setting the scales

The program will automatically set the scale for the X-axis and Y-axis, but you can also set these manually. This is helpful if you want to create a series of plots with the same scale.

- To set the Y-axis manually click Format > Y-axis.
- To set the X-axis manually click Format > X-axis.

STEP 5: SAVE THE ANALYSIS

Once you've run a meta-regression you can save the predictive model as shown in Figure 199.

- Click File > Save regression file as … [A]
- This will save the regression template with an extension of .cmr.

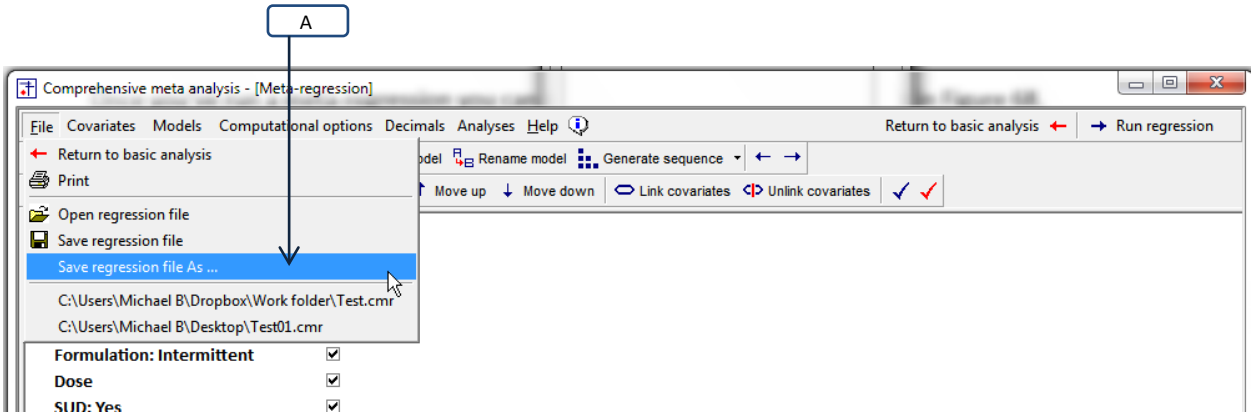


Figure 199 | Save analysis

The .cmr file saves the instructions for the analysis, NOT the data. By analogy, programs such as SPSS™, SAS™, and stata™ allow you to save a set of commands in one file and the data in another file. The commands can then be applied to any data file that has the same variables.

- The .cmr file, saved here, is analogous to the command file in the other programs.
- The .cma file, saved from the data-entry screen, is analogous to the data file in the other programs.

The .cmr file saves the following

- The list of covariates
- The list of models
- The check-boxes for each model
- The sets
- The model names

In another session you can open a data file on the main data-entry screen. Then, return to the regression module and click File > Open file to open the .cmr file and re-run the analysis.

The .cmr file can be used with the same dataset that was used to create it, or with another dataset that includes the same variables. For example,

- You may return to the data-entry screen and add new studies
- You may return to the main analysis screen and edit the study filters
- You may be working with an entirely different data set that has the same variables as the first one.

In any of these cases, navigate to the regression module and click File > Open to open the .cmr file.

When you open a .cmr file the program simply recreates the main MR screen as though you had entered it manually. The .cmr file does *not* save the statistical settings that were in place when the file was created. These include the method employed to estimate T^2, the use of *Z* or Knapp-Hartung, the confidence level, the choice of a one-sided or two-sided test.

STEP 6: EXPORT THE RESULTS

The program offers two options for exporting the results of any analysis.

- Export the results to Excel™. Then, you can perform additional computations within Excel™, and/or format the results and copy them as a table to other programs

- Copy the results to the clipboard as a picture. Then, paste this picture into Word™ or any other program.

Figure 200 shows an example for the main analysis screen.

- Click [File > Save results as Excel™ file and open] [A]
- Provide a name for the Excel™ file

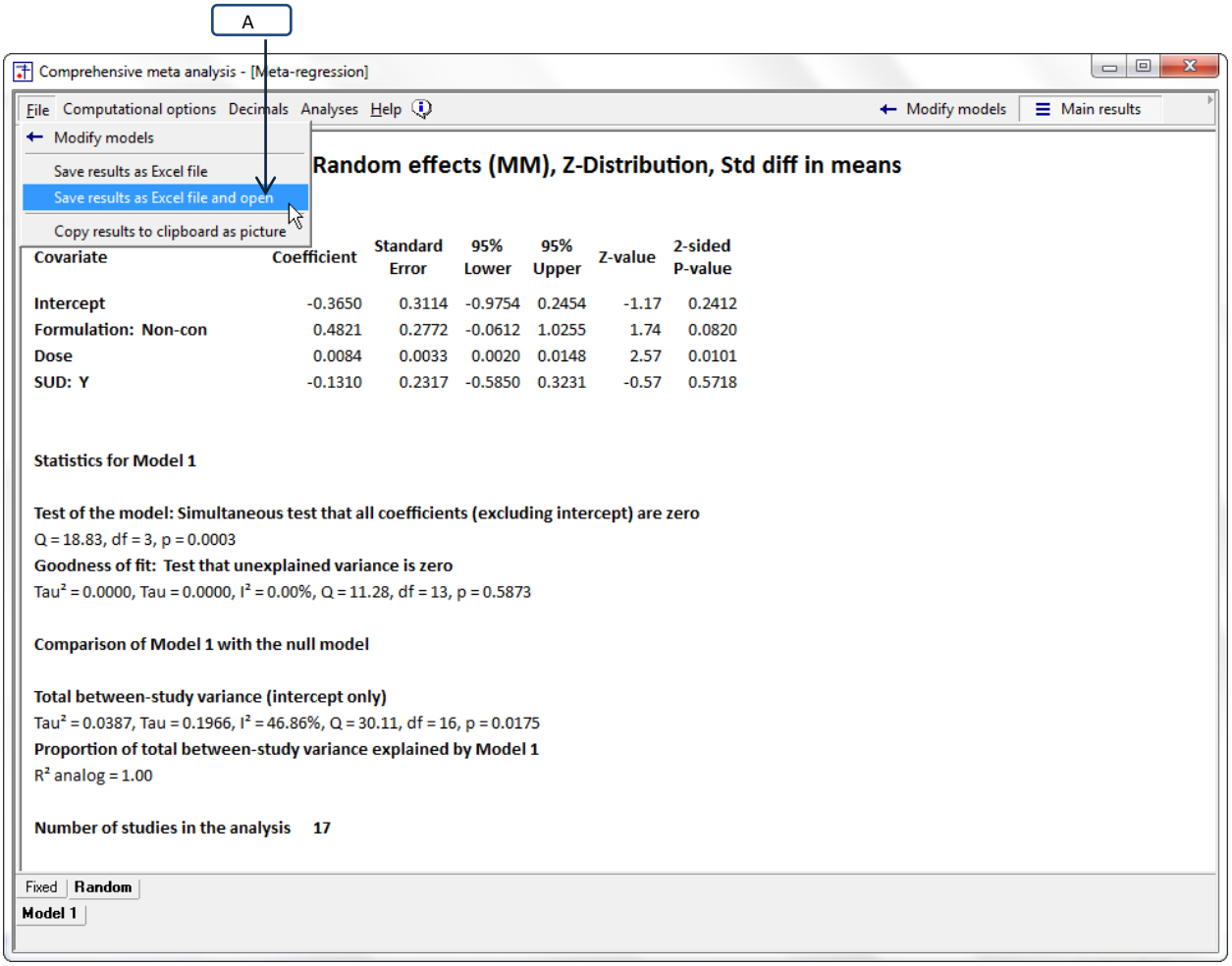
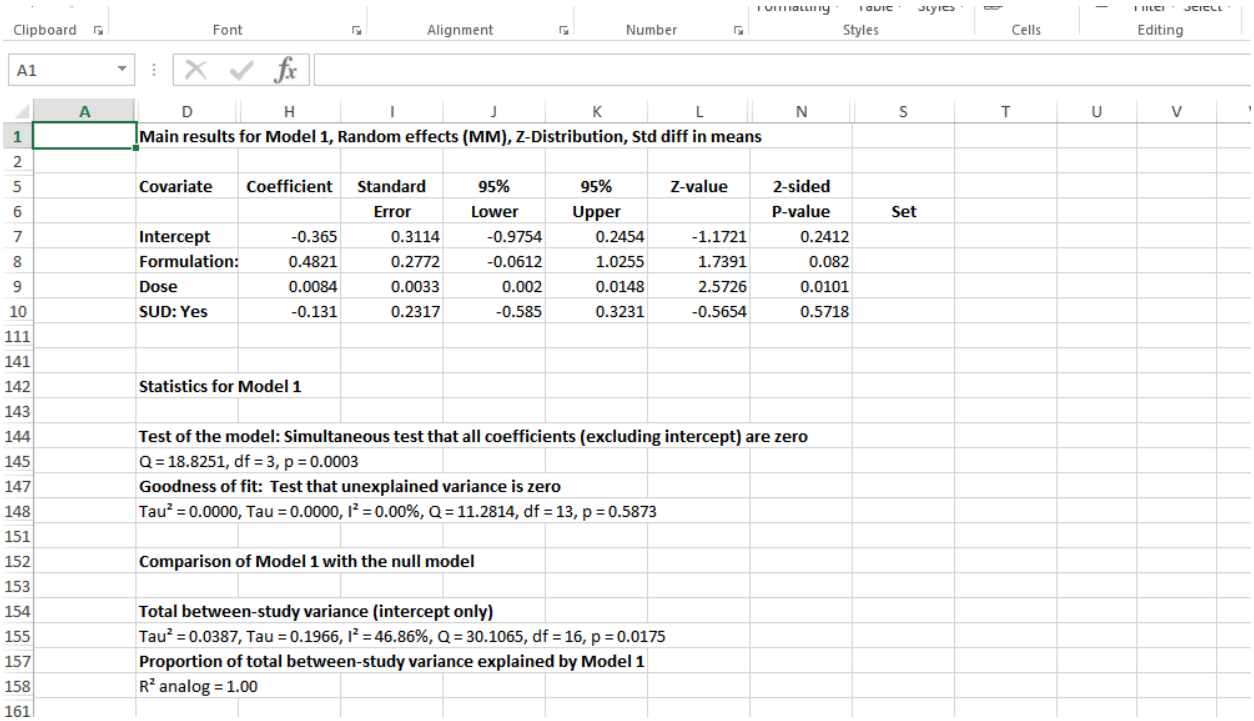


Figure 200 | Export results

The program creates the Excel™ file shown in Figure 201.



	A	D	H	I	J	K	L	N	S	T	U	V
1		**Main results for Model 1, Random effects (MM), Z-Distribution, Std diff in means**										
2												
5		**Covariate**	**Coefficient**	**Standard**	**95%**	**95%**	**Z-value**	**2-sided**				
6				**Error**	**Lower**	**Upper**		**P-value**	**Set**			
7		**Intercept**	-0.365	0.3114	-0.9754	0.2454	-1.1721	0.2412				
8		**Formulation:**	0.4821	0.2772	-0.0612	1.0255	1.7391	0.082				
9		**Dose**	0.0084	0.0033	0.002	0.0148	2.5726	0.0101				
10		**SUD: Yes**	-0.131	0.2317	-0.585	0.3231	-0.5654	0.5718				
111												
141												
142		**Statistics for Model 1**										
143												
144		**Test of the model: Simultaneous test that all coefficients (excluding intercept) are zero**										
145		Q = 18.8251, df = 3, p = 0.0003										
147		**Goodness of fit: Test that unexplained variance is zero**										
148		Tau^2 = 0.0000, Tau = 0.0000, I^2 = 0.00%, Q = 11.2814, df = 13, p = 0.5873										
151												
152		**Comparison of Model 1 with the null model**										
153												
154		**Total between-study variance (intercept only)**										
155		Tau^2 = 0.0387, Tau = 0.1966, I^2 = 46.86%, Q = 30.1065, df = 16, p = 0.0175										
157		**Proportion of total between-study variance explained by Model 1**										
158		R^2 analog = 1.00										
161												

Figure 201 | Export results

The same idea applies to any screen that displays results.

THE RANDOM-EFFECTS RESULTS

Table 1

Section	Function	Covariates	Weights
D E	Random-effects estimates	Yes	V+T2
F	Variance not explained by model	Yes	V
G	Original variance	No	V
H	Proportion of variance explained (Based on F and G)		

Sections D and E report statistics for an analysis that employs random-effects weights and includes the covariates. This provides a test of the model, and is also the analysis used in the table at the top of the screen.

Section F reports statistics for an analysis that includes the covariates but assigns weights based on V (within-study error variance). This provides a goodness-of-fit test. Specifically, we use this analysis to estimate the residual T2, the variance not explained by the covariates.

Section G reports statistics for an analysis that does not include covariates and assigns weights based on V. This allows us to estimate the original T2, the total amount of variance (see appendix)

Section [H] is based on the analyses in sections [F] and [G]. Section [F] gives us the variance that cannot be explained by the covariates, while section [G] gives us the total variance. We can use these to compute the ratio of Explained/Total, which is presented in section [] (See analysis 5 – Dose and SUD as the covariates)

In Analysis-3 we included only Dose, and this allowed us to assess the impact of Dose while ignoring the potential confound with SUD.

In Analysis-4 we included only SUD, and this allowed us to assess the impact of SUD while ignoring the potential confound with Dose.

Now, we will run an analysis with both Dose and SUD as covariates. This will allow us to assess the unique impact of each covariate, as well as the combined impact of the two.

The results are shown in Figure 15 AND PLOTTED IN FIGURE 16.

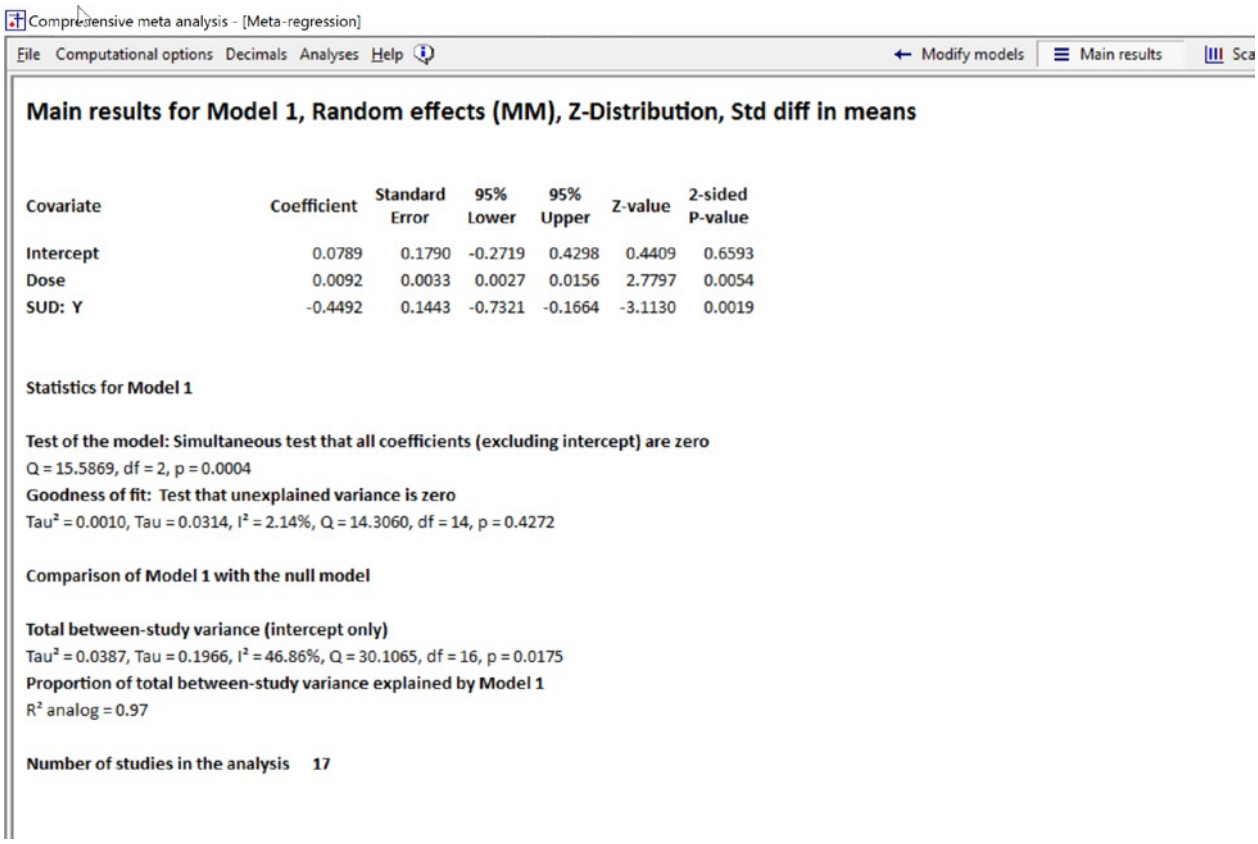


Figure 15

IS THIS REDUNDANT WITH EARLIER ?
BE SURE TO INCLUDE SAMPLE REPORT ABOVE

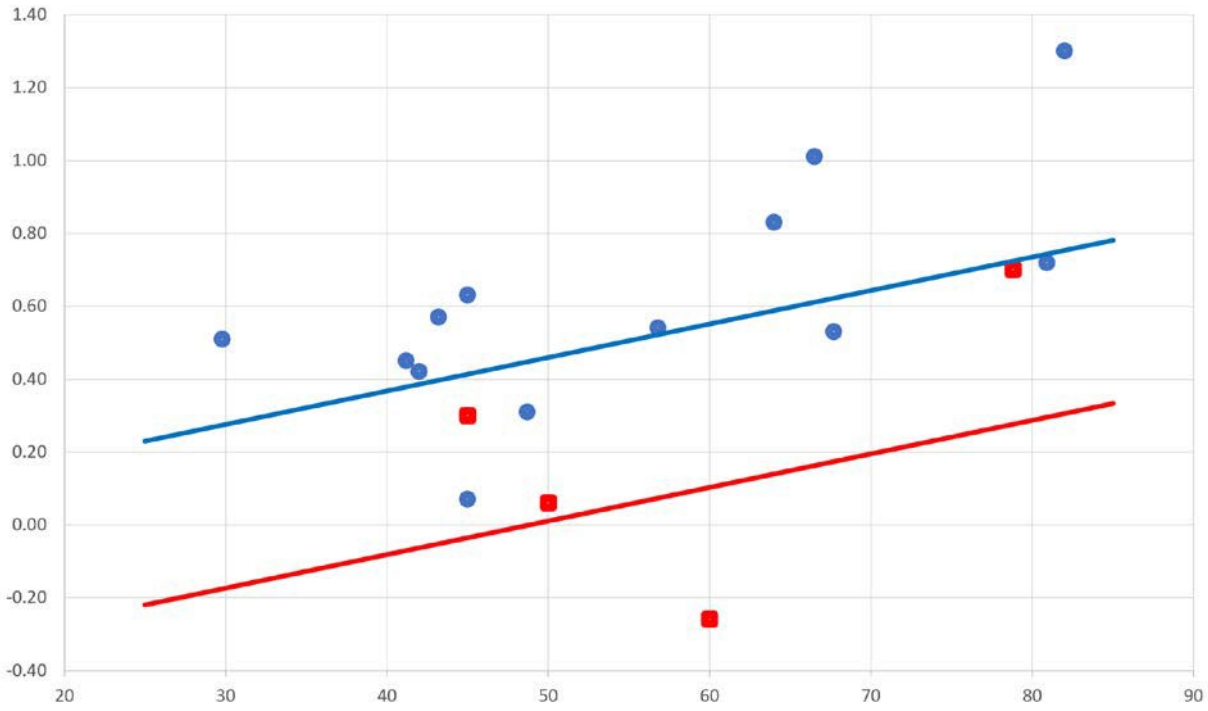


Figure 16

The covariates

Here, we address section [A] in Figure 15.

Relationship between Dose and effect size when SUD is held constant

The line for Dose displays the relationship between Dose and effect size, controlling for SUD. The coefficient for Dose is 0.0092. The fact that the coefficient is positive tells us that a higher dose is associated with a higher effect size. Specifically, a one unit increase in dose corresponds to an increase of 0.0092 in effect size. To make this more intuitive, we could multiply the coefficient by 50, and say that a 50-point increase in dose corresponds to an increase of around 0.46 in the d-value.

The coefficient reported here is an estimate of the parameter (the population value), and the other statistics on this line speak to the precision of the estimate. The standard error is 0.0044. If we assume that the true coefficient usually (in 95% of cases) falls within 1.96 standard error of the estimate, then it probably falls in the range of 0.0027 to 0.0156, which is displayed as the 95% confidence interval. At the lower end, a one unit increase in dose would be associated with an increase of 0.1350 in the d-value (and a 50-point increase in dose would be associated with an increase of around 4 points). At the upper end, a one-unit increase in dose would be associated with an increase of 0.0156 in the d-value (and a 50-point increase in dose would be associated with an increase of around 78 points in the d-value).

Finally, we can test the null hypothesis that the true value of the coefficient is zero (that there is no relationship between Dose and effect size when SUD is held constant). The 95% confidence interval (0.0027 to 0.0156) does not include zero. Similarly, the Z-value (computed as the coefficient divided by its standard error) is 2.7797, and the corresponding *p*-value is 0.0054 (which is less than the criterion alpha of 0.05). In either case, we reject the null hypothesis that the true value of the coefficient is zero, and

conclude that there probably is a positive relationship between dose and effect size when SUD is held constant.

Relationship between SUD and effect size when Dose is held constant

The line for SUD| Y displays the relationship between SUD and effect size, when Dose is held constant. In Chapter _____ we discuss how to interpret coefficients for categorical variables, but for now we can simply say that the coefficient gives us the mean difference in effect size for studies that enrolled SUD patients as compared with those that excluded them, assuming the same value of Dose. The coefficient −0.4492, which tells us that the mean effect size for studies that enrolled these patients is some 45 points lower than for studies that included them.

THE EFFECT SIZES DIDN'T CHANGE MUCH. THE SE GOT SMALLER

The coefficient reported here is an estimate of the parameter (the population value), and the other statistics on this line speak to the precision of this estimate. The standard error is 0.1443. If we assume that the true coefficient usually (in 95% of cases) falls with 1.96 standard error of the estimate, then it probably falls in the range of −0.7321 to −0.1664, which is displayed as the 95% confidence interval. For a given dose, the mean effect size for the SUD studies could be as little as 16 points lower than the non-SUD studies, or as much as 73 points lower.

Finally, we can test the null hypothesis that the true value of the coefficient is zero (that there is no relationship between SUD and effect size) when Dose is held constant. The 95% confidence interval (−0.7321 to −0.1664) does not include zero. Similarly, the Z-value (computed as the coefficient divided by its standard error) is −3.1130 and the corresponding p-value is 0.0019 (which is less than the criterion alpha of 0.05). In either case, we reject the null null hypothesis that the true value of the coefficient is zero, and conclude that (with Dose held constant) the mean score for SUD studies is probably lower than that for non-SUD studies.

The model

The section labeled "Statistics for Model 1" includes rows labelled "Test of the model" and "Goodness of fit". The first asks if the model is able to explain *any* of the variation in effect size – is the variance of effects about the regression line smaller than the variance of effects about the mean? The second asks if the model is able to explain *all* of the variation in effect size – is the variation of effects about the regression line more than we would expect to see based on sampling error alone?

The predictive model

Here, we address section [B] in Figure 15.

Both sections [A] and [B] address the relationship between covariates and the effect size. However, there are two key difference between these two sections.

Where section [A] provides information about the direction and magnitude of the relationship, section [B] only tests the presence or absence of a relationship.

Where section [A] speaks to the impact of each covariate with all other covariates held constant, section [B] speaks to the combined impact of the full set of covariates. The Q-value provides a test of the null hypothesis that the true coefficient for all covariates is zero.

In Analysis-3 and Analysis-4 there was only one covariate, and therefore sections [A] and [B] were testing the same model. That's not the case now. In the current analysis, each line in section [A] addressed the unique impact of one covariate. By contrast, section [B] addresses the combined impact of the two covariates. The Q value is 15.5869, with two degrees of freedom (corresponding to the two covariates) and a p-value of 0.0004. Since the p-value is less than the criterion alpha of 0.05, we reject the null hypothesis and conclude that at least one of the covariates is related to the effect size.

Goodness of fit

This covers sections [C] in Figure 15.

The statistics presented here have the same interpretation as those presented in the prior regressions. In all cases, the statistics describe the dispersion of effects aBOUT THE REGRESSION LINE. IN Analysis-2, the regression line was horizontal – the predicted effect size for all studies was the same. In Analysis-3, the regression line was based on Dose. In Analysis-4, the regression line (actually displayed as two lines) was based on SUD. In the current analysis, the regression line is based on Dose and SUD. As such, it takes the form of two lines as in Figure 16 – there's one regression line that shows the relationship between Dose and effect size for SUD studies, and another that shows the relationship between Dose and effect size for non-SUD studies.

WHEN WE SAY THAT MOST EFFECTS FALL WITH TWO SD, WHAT ABOUT FACT THAT PI LINES ARE NOT PARALELL – SHOULD IT BE AT THE MEAN?

The standard deviation of true effects, T, is 0.0314. If we assume that most true effects will fall within two standard deviations of the regression line, then most will fall within 0.06 of the regression line. For a study that enrolls SUD patients and uses a Dose of 30 units, the predicted effect size is ____. The true effect size for that study probably falls in the range of _____ to _____.

The variance of true effects, T2, is 0.0010. This is simply the standard deviation, squared. This is a less intuitive metric than T, but it has some useful statistical properties.

The Q statistic may provide a test of the null hypothesis that the true effect size for all studies is the predicted value. This could also be framed as "The true effect size for all studies lies precisely on the regression line."

The Q-value is 14.3060 with 14 degrees of freedom and a p-value of of 0.4272. When we test for heterogeneity we typically use a criterion alpha of 0.10. On this basis we would not reject the null. The dispersion of effects about the regression line could be due entirely to sampling error. Another way to say this, is that the actual value of T (AND T2) could be zero.

Finally, I2 is reported as 2.14%. As always, I2 does not tell us how much the effects vary. Rather, it provides some context for the variance of observed effects. While the plot shows some variation in observed effects about the regression lines, only about 2% of this reflects variation on true effects. If we

could somehow plot the true effect size (rather than the observed effect size) for each study, the effects would tend to fall much closer to the regression line than they do in Figure 16.

Comparison of Model 1 with the null model

When we perform a regression analysis in a primary study we often report a statistic called R2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T20 is the variance of true effects about the mean, and T21 is the variance of true effects about the regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

Total between-study variance (intercept only)

When we perform a regression analysis in a primary study we often report a statistic called R2, which is the proportion of variance explained by the covariates. We can report an analogous statistic here. If T20 is the variance of true effects about the mean, and T21 is the variance of true effects about the regression line with covariates, the difference between them gives us the amount of variance explained by the covariates. This value, divided by the initial variance, gives us the proportion of variance explained. This is addressed in sections [D] and [E].

In section [D] we saw that with no covariates in the model, the variance of true effects is 0.0387. This is the same value reported in Analysis-1 and Analysis-2, and it will serve as the baseline value for computing R2. In section [C] we saw that the variance of true effects about the regression line based on Dose and SUD was 0.0010. The difference,

$$T^2_{Explained} = T^2_{Total} - T^2_{Residual} = 0.0387 - 0.0010 = 0.0377 \tag{1.125}$$

We can then estimate the proportion of variance explained by SUD as

$$R2 = \frac{T^2_{Explained}}{T^2_{Total}} = \frac{0.0377}{0.0387} = 0.97 \tag{1.126}$$

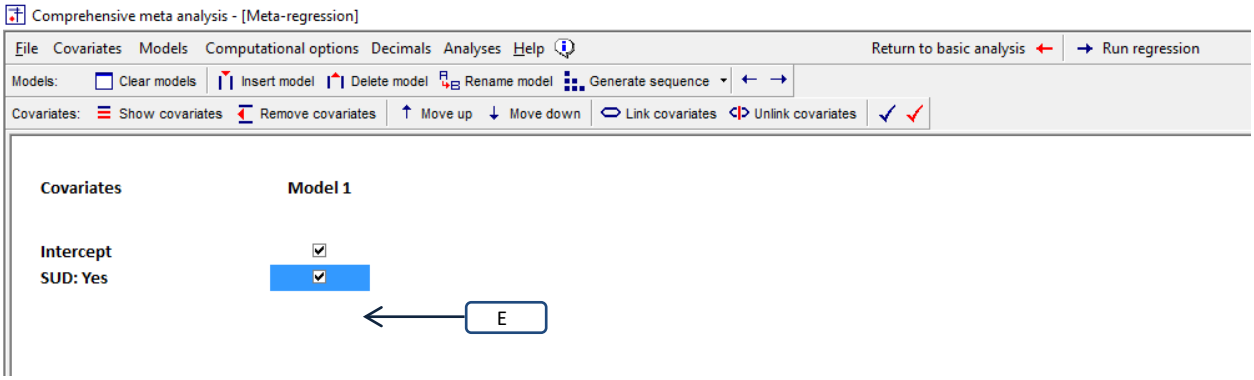


Figure 202

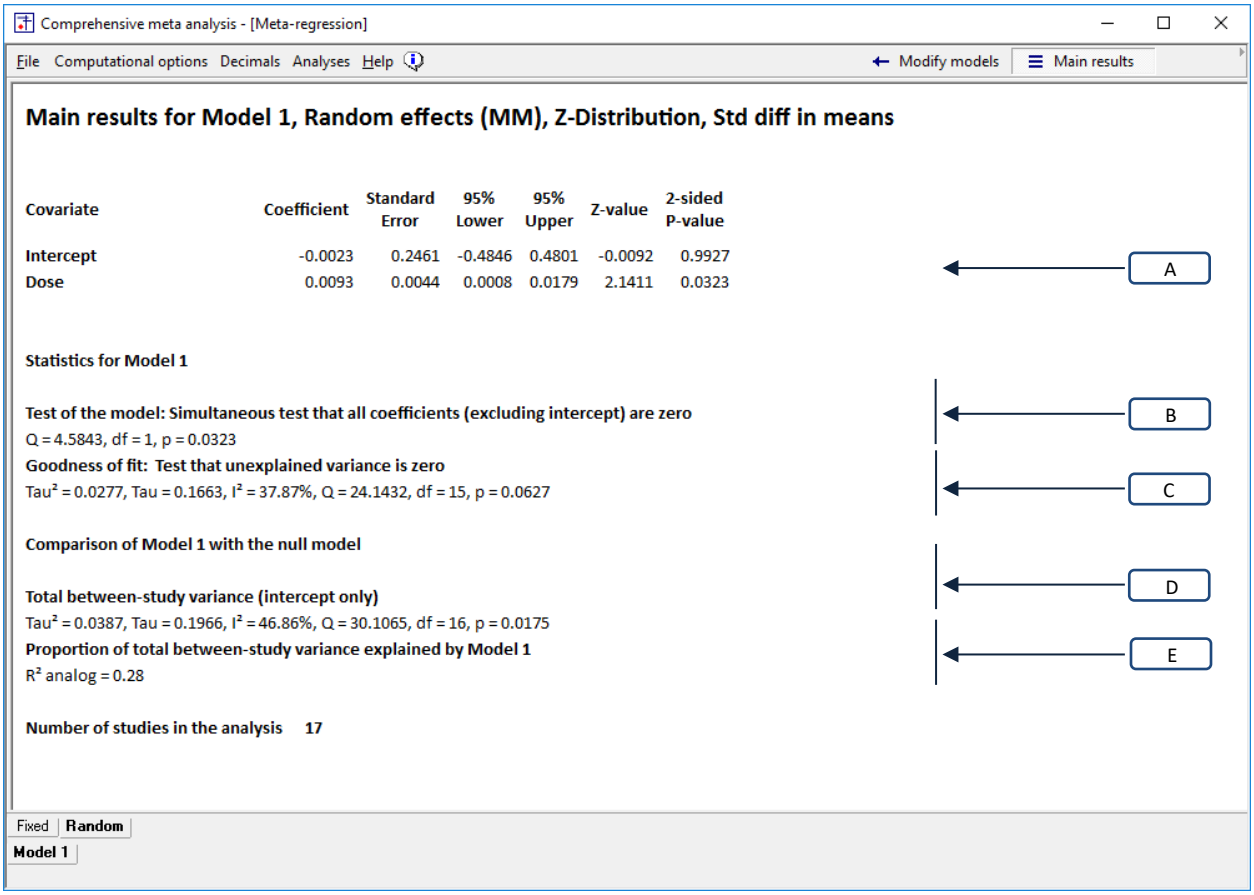


Figure 203 | Regression | Dose | Main results | Random-effects

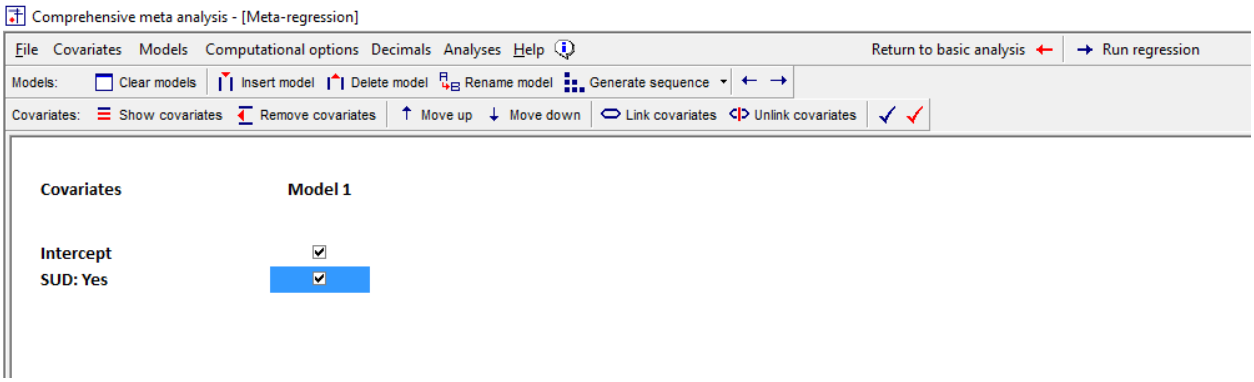


Figure 204

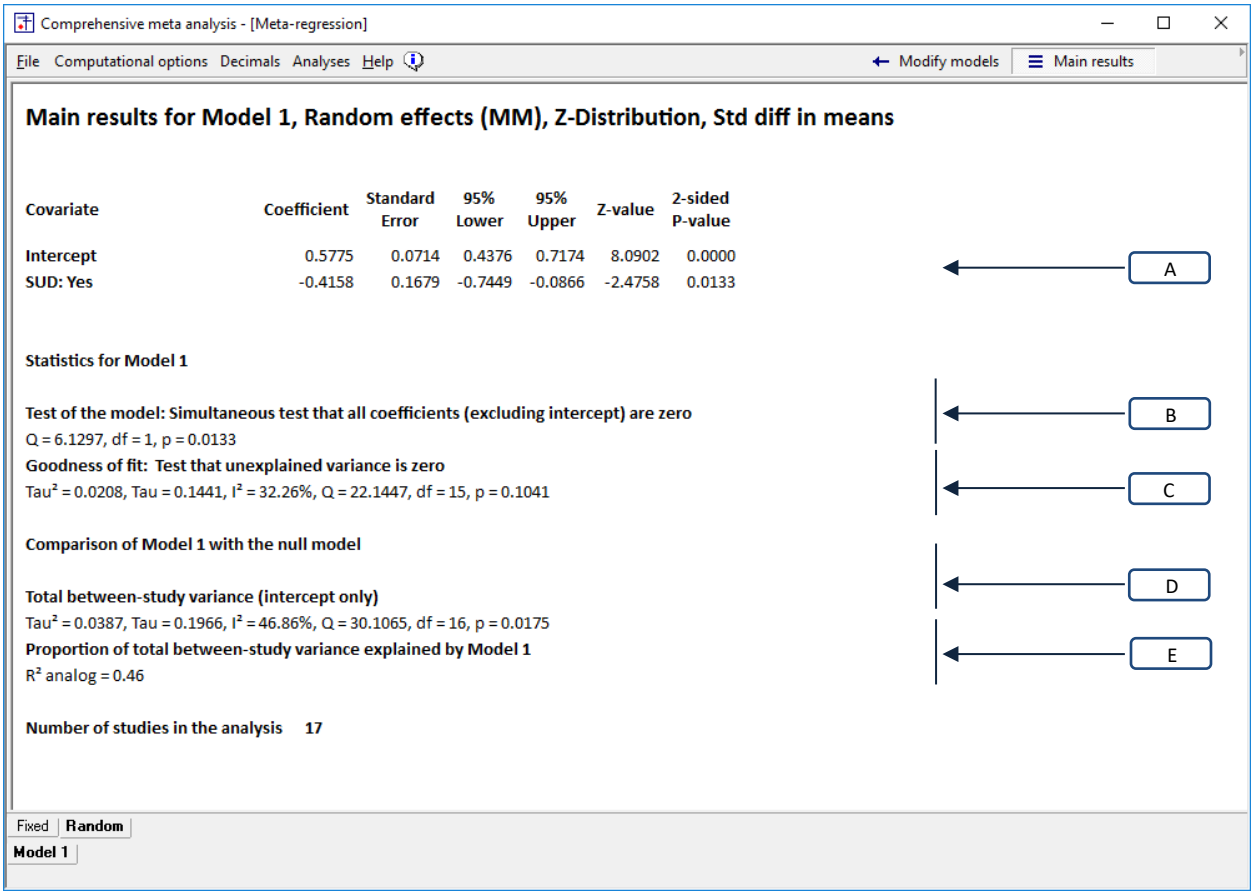


Figure 205 | Regression | Dose | Main results | Random-effects

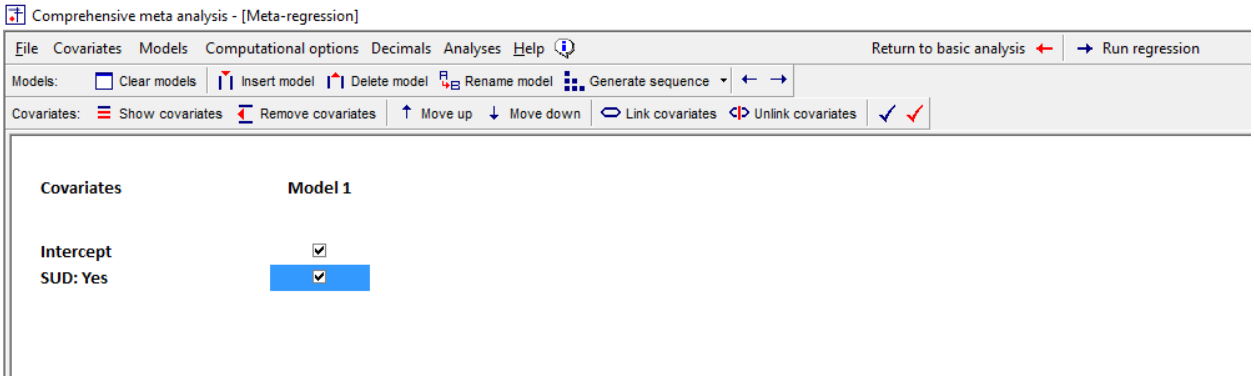


Figure 206

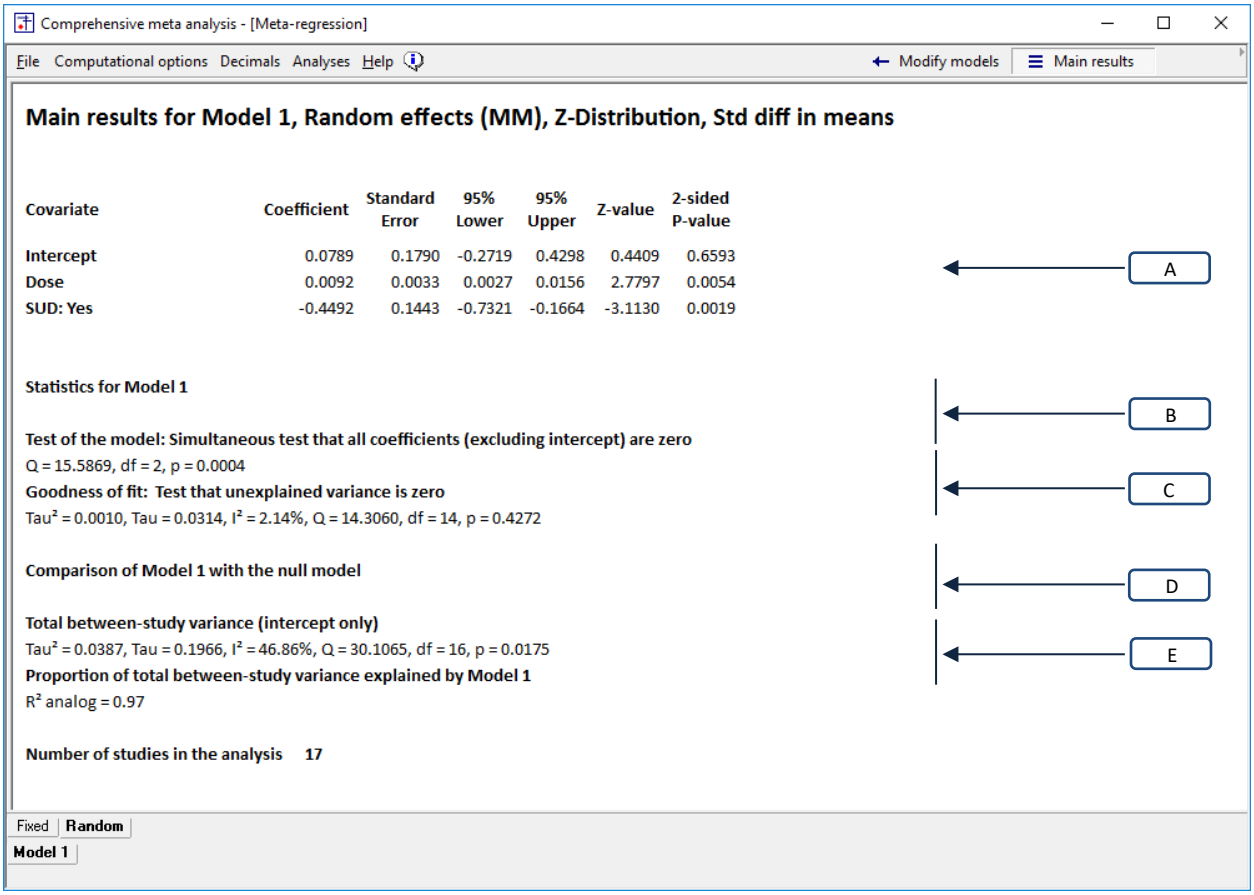


Figure 207 | Regression | Dose | Main results | Random-effects

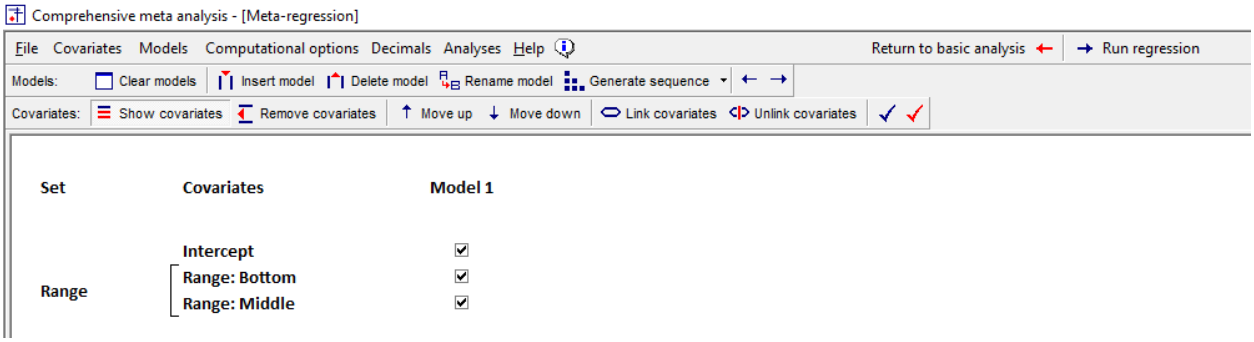


Figure 208

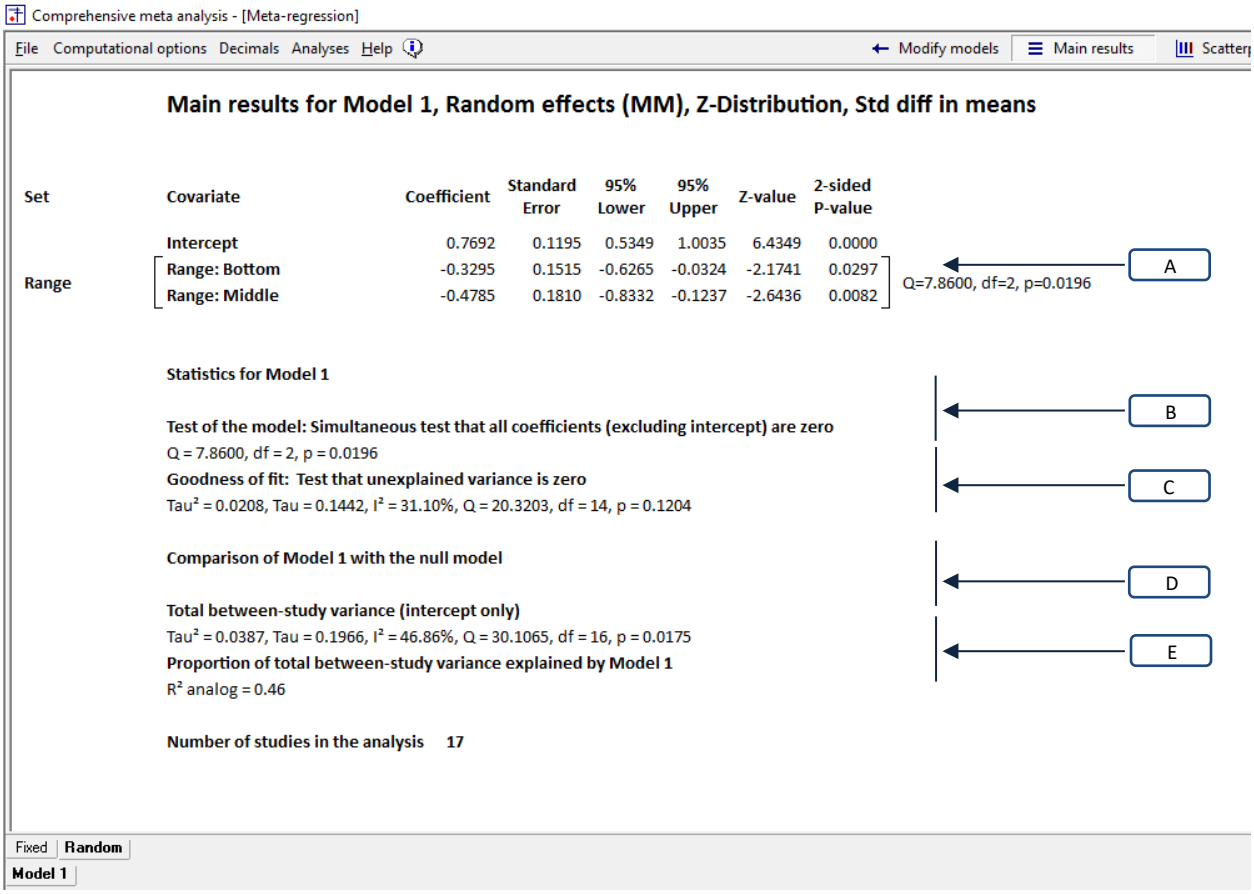


Figure 209 | Regression | Dose | Main results | Random-effects

Summary

We used meta-regression to examine the relationship Dose (as a categorical variable) and effect size. Studies were classified as Low, Moderate or High dose. The analysis treats these groups as categorical and ignores the possibility of any ordinal relationship.

The mean effect size for the Low-Dose studies is ___, for the Moderate-Dose studies is ____, and for the High-Dose studies is _____. The Q-value for the model is 7.8600 with 2 degrees of freedom and a corresponding p-value of 0.0196, so we conclude that the mean effect size varies by group. [B]

The difference between the High-Dose studies and the Low-Dose studies is 0.3295 with a 95% confidence interval of 0.6265 to 0.0324 (Z=2.1741, p=0.0297). The difference between the High-Dose studies and the Moderate-Dose studies is 0.4785 with a 95% confidence interval of 0.8332 to 0.1237 (Z=2.6436, p=0.0082). [A]

The variance of true effects about the subgroup means (T^2) is 0.0208, and the standard deviation of true effects about the regression line (T) is 0.1442. The I^2 statistic is 31.10%, which tells us that about 31% of the observed variance about the subgroup means reflects variation in true effects rather than sampling error. The test for heterogeneity yields a Q-value of 20.3203 with 14 degrees of freedom and a corresponding p-value of 0.1204. There is no evidence that the true effects vary across studies within subgroups. [C]

The R^2 analog is 0.46, which means that the model is able to explain some 46% of the variance in true effects. [E]

The relationship between range and effect size is observational. The fact that the treatment-effect is higher in some subgroups than in others could be due to the difference in dose, but could be due to other factors as well.

Suppose we wanted to see if Dose was related to effect size as is, and also when we partial SUD – would run two analyses

Suppose we wanted to see if Range is related to effect size, and also do each of the pairwise comparisons

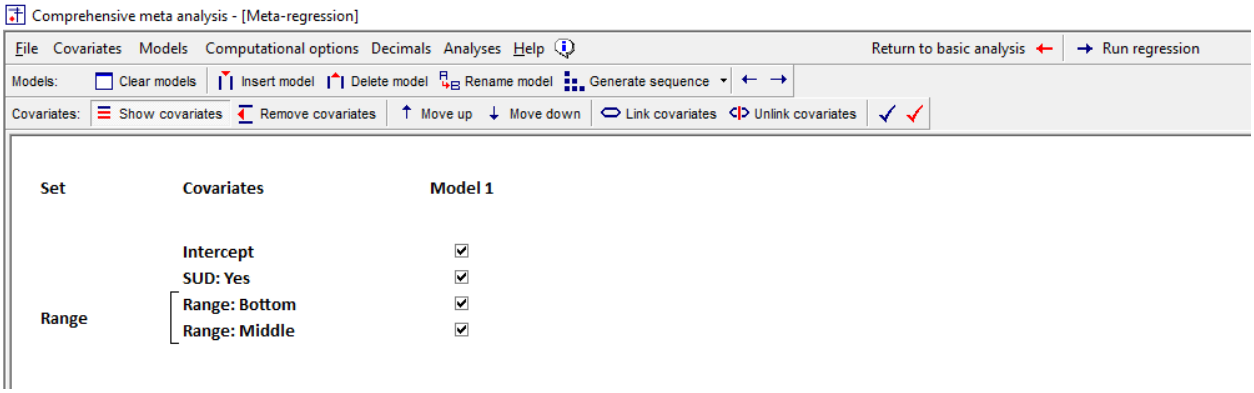


Figure 210

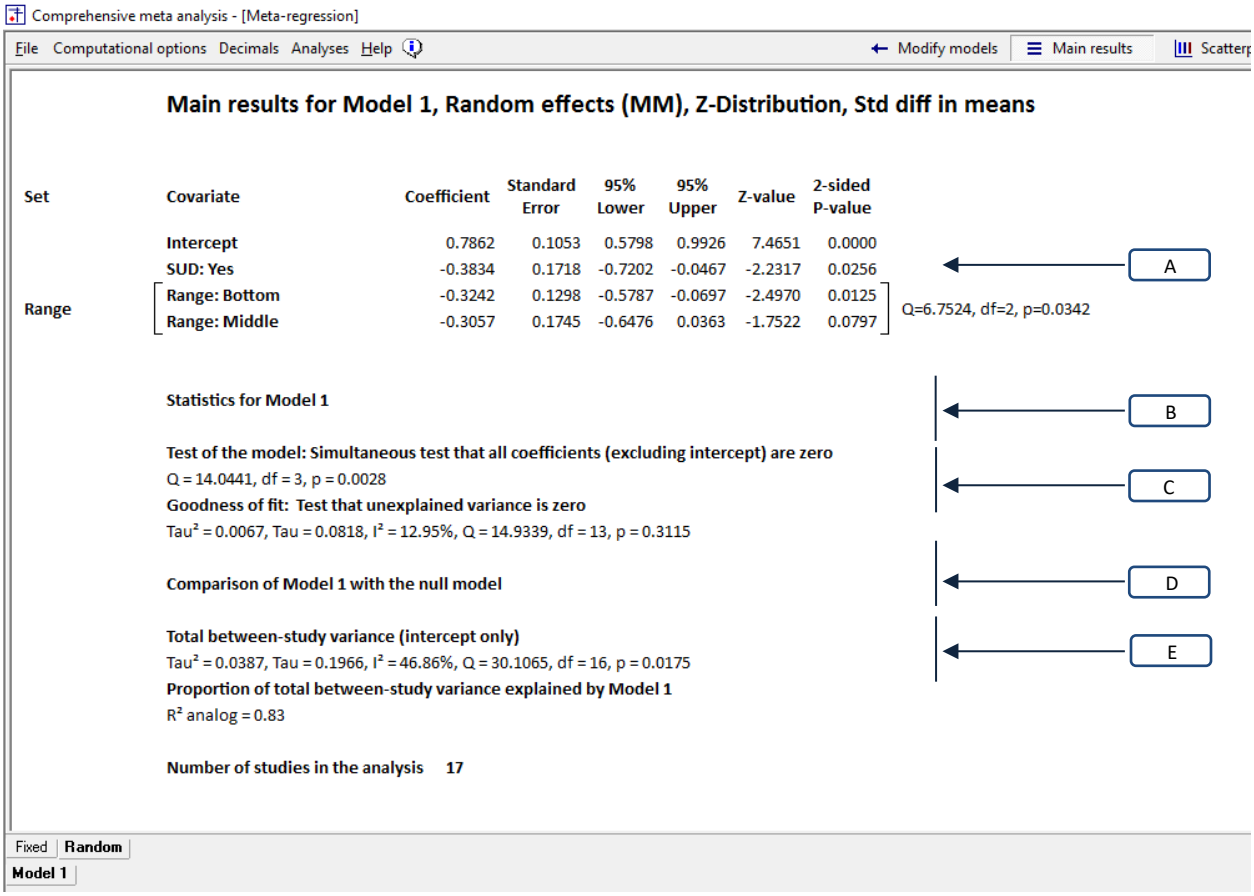


Figure 211 | Regression | Dose | Main results | Random-effects

Summary

We used meta-regression to examine the relationship Dose (as a categorical variable) and effect size, controlling for SUD. Studies were classified as Low, Moderate or High dose. The analysis treats these groups as categorical and ignores the possibility of any ordinal relationship.

The variables SUD and Range (together) are able to explain at least some of the variance in effect size. The test of the model yields a Q-value of 14.0441, df=2, p=0.0028.

Additionally, each of the factors explains a unique aspect of the variance.

With Range partialled, the effect size is lower for studies that enrolled SUD patients as compared with studies that excluded these patients. The difference is −0.3834 with a confidence interval of −0.7202 to −0.0467 (Z=−2.2317, p=0.0256).

With SUD partialled, the effect size differs among the three subgroups. The test for differences among all subgroups yields a Q = 6.2574 with 2 degrees of freedom and p=0.0342. Pairwise comparisons among the subgroups shows that the effect size is higher in the High-Dose group than either of the others (0.3242 higher than the Low-Dose, with a 95% CI of 0.5787 to 0.0697, Z=2.4970, p=0.0125; and .3057 higher than the Moderate-Dose, with a 95% CI% of 0.6476 to −0.0363, Z=1.7522, p=0.0797).

The variance of true effects about the regression line (T^2) is 0.0067, and the standard deviation of true effects about the regression line (T) is 0.0818. The I^2 statistic is 12.95%, which tells us that about 13% of the observed variance about the regression line reflects variation in true effects rather than sampling error. The test for heterogeneity yields a Q-value of 14.9339 with 13 degrees of freedom and a corresponding p-value of 0.3115. There is no evidence that the true effects vary about the regression. [C]

The R^2 analog is 0.83, which means that the model is able to explain some 83% of the variance in true effects. [E]

The relationship between the covariates and effect size is observational. The fact that the treatment-effect is higher in some subgroups than in others could be due to SUD and Range, but could also be due to other factors that are confounded with SUD and Range.

By comparing these two analyses we can appreciate how the regression operates. In the first one, the predicted effect size is lowest for the Moderate-Dose studies, but in the second analysis the predicted effect size for this subgroup has increased. That's because the studies that enrolled SUD patients tend to fall in this subgroup

Suppose we wanted to see if Dose was related to effect size as is, and also when we partial SUD – would run two analyses

Suppose we wanted to see if Range is related to effect size, and also do each of the pairwise comparisons

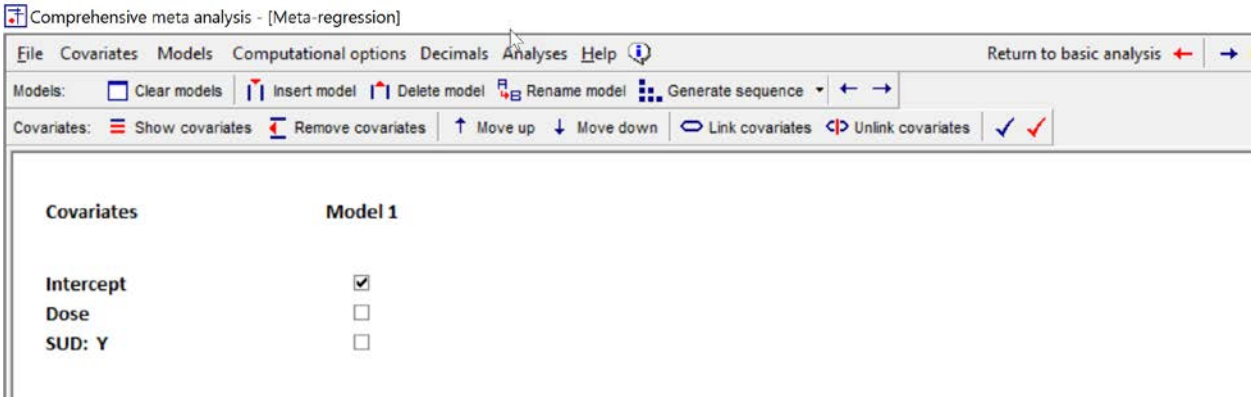
WORKING WITH MULTIPLE MODELS

Immediately above, we showed how we might want to run a series of analyses to develop a more complete understanding of the relationships among the covariates. In our example we included Analysis-1 and Analysis-2 to explain to provide a foundation for explaining the statistics. In a real analysis, one would probably want to exclude these, and run the following analyses.

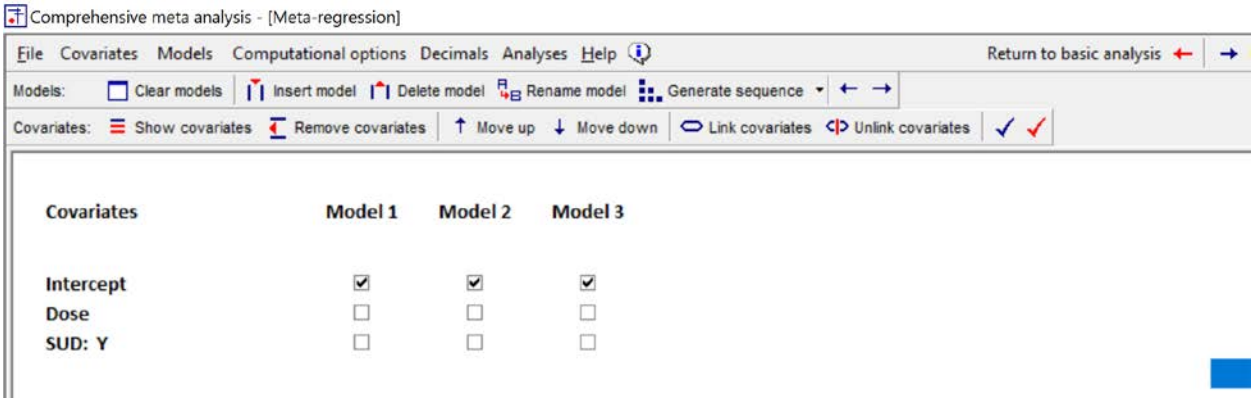
- Analysis-3 looks at the impact of Dose
- Analysis-4 looks at the impact of SUD
- Analysis-5 looks at the impact of Dose and SUD

While it's possible to run each analysis in turn, as above, it would be more convenient to run all three analysis at once, and have access to all three sets of results at the same time. CMA allows you to do so, by creating a series of models.

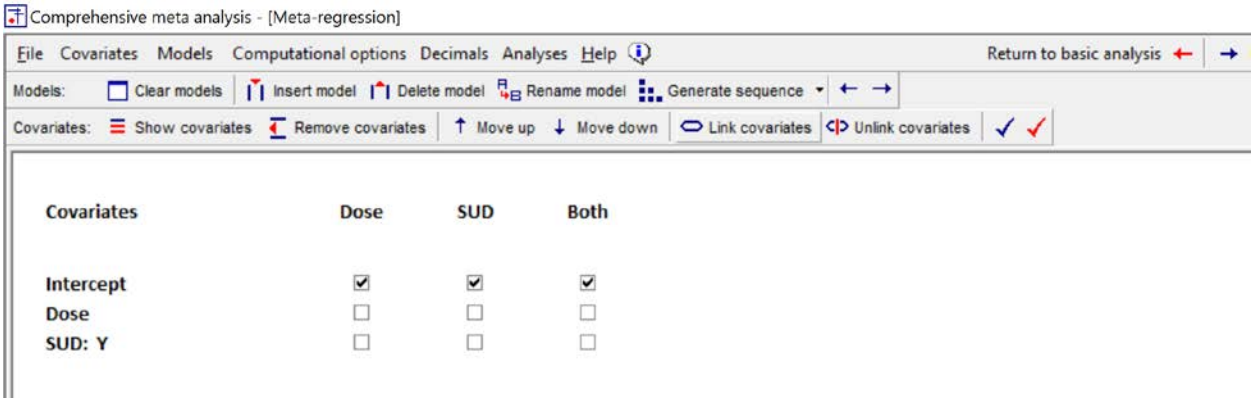
Follow the standard procedure to add Dose and SUD|Y to the main screen. Initially, the program creates a model called Model 1.



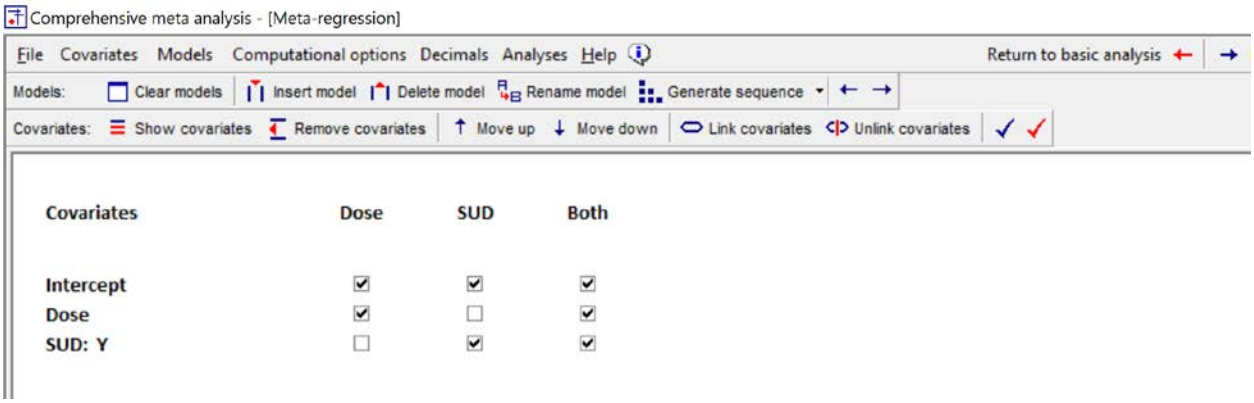
Click "Insert Model" two times. The program adds two additional models, called Model 2 and Model 3.



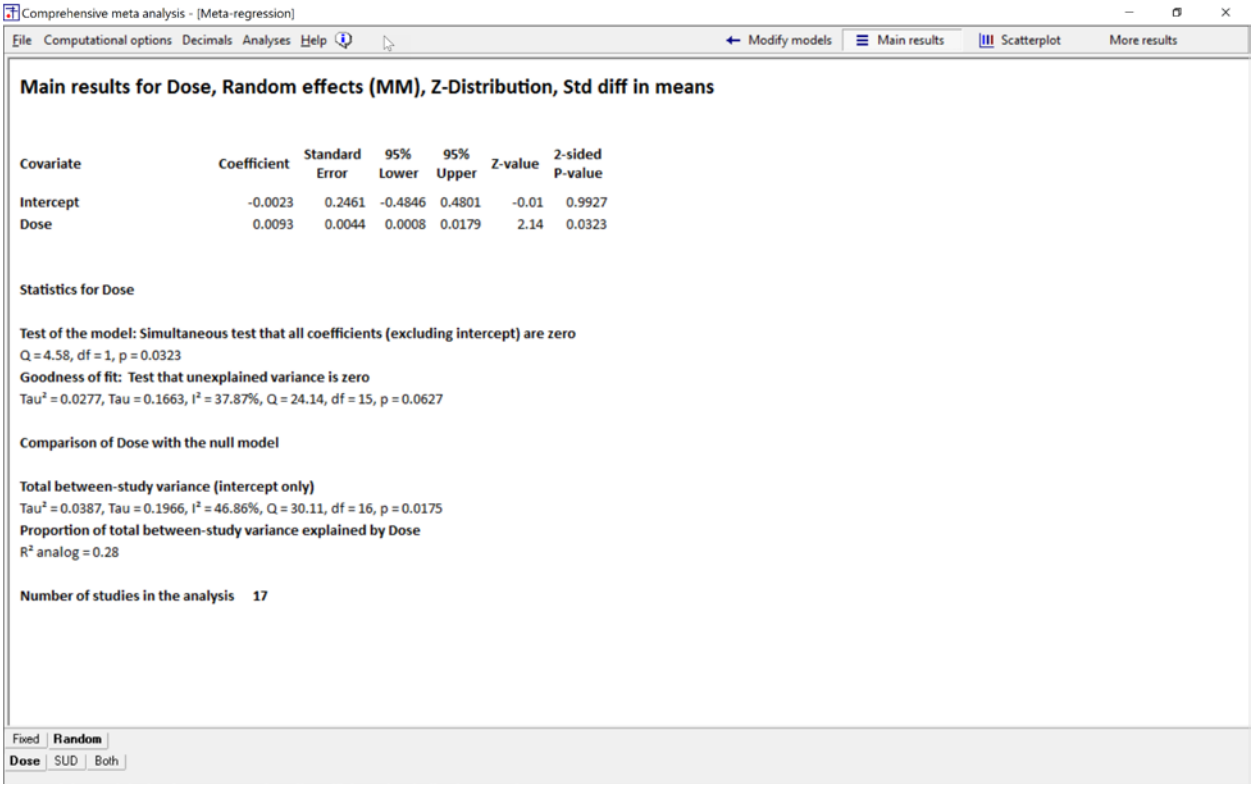
Right-click on each Model name, and rename the models to Dose, SUD, and Both



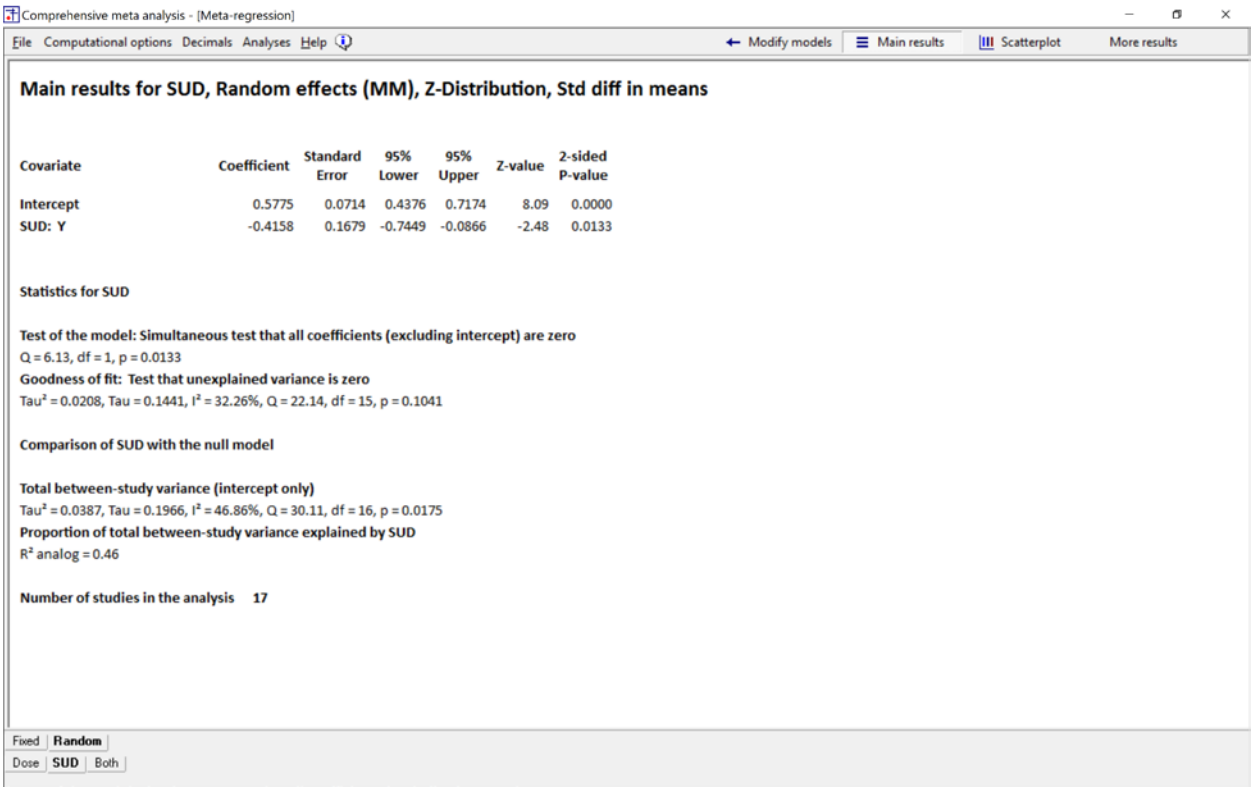
Tick the appropriate boxes for each model



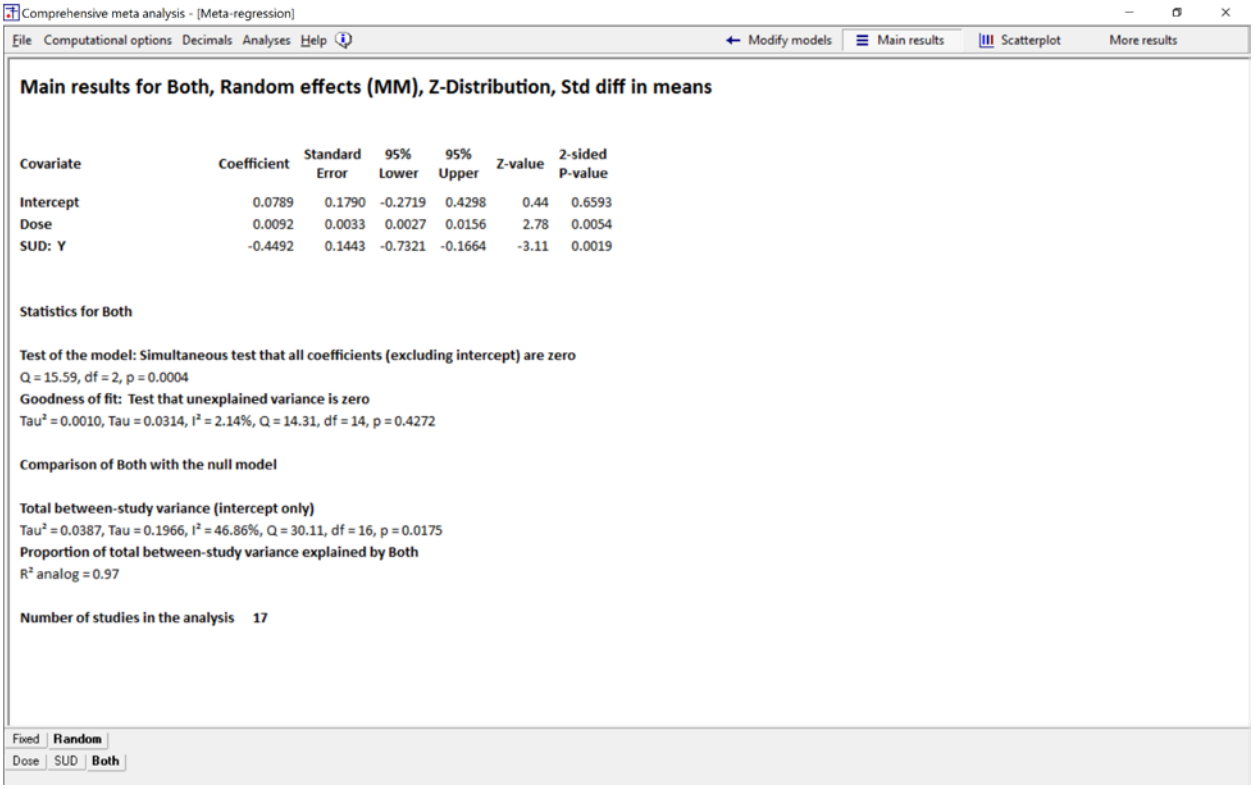
Run the analysis



When you run the analysis, there are three tabs at the bottom of the screen, corresponding to the three models. Click on Dose and the program displays this screen.



Click on SUD and the program displays this screen



Click Both and the program displays this screen.

Similarly, if you select the plot (for example) the program will display the plot corresponding to either Dose, SUD, or Both.

To be clear, the names Dose, SUD, and Both were selected to describe the three models, but have no impact on the actual model. Rather, the model is defined by the boxes selected in each case.

Appendix – The standardized mean difference, d

In the ADHD example, the effect size is the standardized mean difference, d. Meta-regression can be used with any effect size index, and all the key points in this volume apply regardless of what the effect size index might be.

For purposes of reading this book, readers who are not familiar with the standardized mean difference can think of it simply as a difference in means on a standardized scale. That is, it's a scale where a difference of 0 points indicates that the two groups had the same mean; a difference of +1 indicates that one group's mean (say, the treated group) was one standard deviation higher than the other; and a difference of -1 indicates that the other group's mean (say, the control group) was one standard deviation higher than the other. In the social science, it's common to see d-values of 0.2, 0.5, and 0.8, respectively, referred to as small, moderate, and large effects.

The standardized mean difference, d, is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \qquad (1.127)$$

where μ1 and μ2 are the population means, and σ is the common within-population standard deviation.

In the ADHD analysis, the sample estimate d is given by

$$d = \frac{M_1 - M_2}{s} \qquad (1.128)$$

where M1 and M2 are the sample, and s is the estimate of the common within-group standard deviation.

For example, suppose that some studies tested patients on a scale where scores ranged from 0 to 100, and the standard deviation was 20 points. A 10-point difference between groups would yield a d-value of

$$d = \frac{M_1 - M_2}{s} = \frac{10}{20} = 0.50 \qquad (1.129)$$

Similarly, suppose that some studies tested patients on a scale where scores ranged from 0 to 1000, and the standard deviation was 200 points. A 100-point difference between groups would yield a d-value of

$$d = \frac{M_1 - M_2}{s} = \frac{100}{200} = 0.50 \qquad (1.130)$$

The difference of 10 points in the first example, and the difference of 100 points in the second example, both carry the same clinical meaning – in both cases, the treatment improved functioning by one-half a

standard deviation. By converting both scales into the same units, we can see that they are both the same, and can include both studies in the same meta-analysis.

Dose. Was the effect size related to the dose of methylphenidate?
Duration. Was effect size related to the duration of treatment?
Formulation. Some studies employed a continuous-release mechanism while others employed a bi-phasic mechanism, where the drug is administered intermittently. Was effect size related to this mechanism?
SUD. Some studies exclude patients with a diagnosis of substance use disorder (SUD), while others allowed them to enroll in the study. Was this related to effect size?

1 APPENDIX I – STATISTICS FOR HETEROGENEITY

When researchers discuss heterogeneity, they typically report an array of statistics which may include *Q*, *df*, *p*, I^2, *I*, T^2, and *T*. Figure 212 shows the relationships among these statistics.

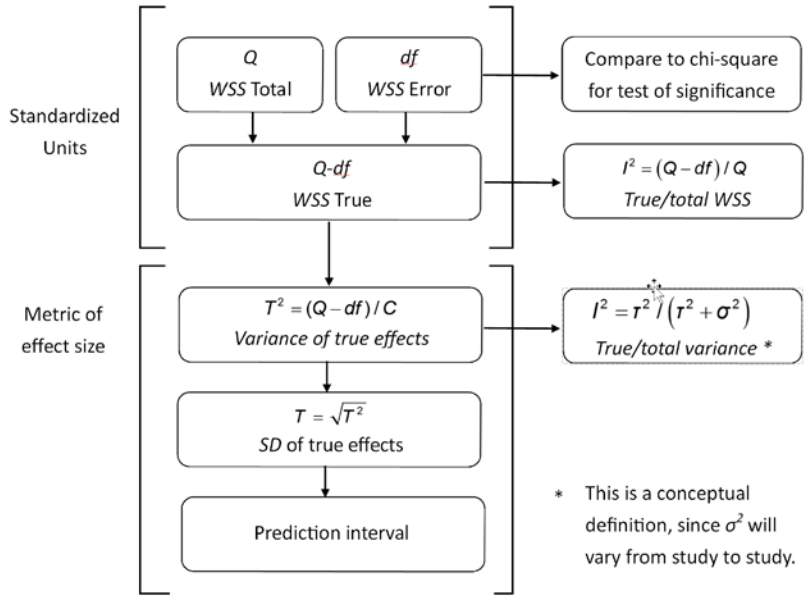


Figure 212

Computing *Q* and *df*

The *Q*-value refers to the distribution of *observed* effects. The *Q*-value is the sum of squared deviations of all effects about the mean, on a standardized scale. Concretely,

$$Q = \sum \left(\frac{X_i - M}{SE_{X_i}} \right)^2$$

where X_i is the effect size in the i^{th} study, *M* is the mean effect size using fixed-effect weights, and SE_{Xi} is the standard error of the i^{th} study. On this scale, the value of *Q* we would expect to see based on sampling error alone is equal to *df* (the degrees of freedom) which is the number of studies minus 1.

These two numbers (*Q* and *df*) serve as the basis for all the other statistics, as follows.

Computing a p-value

If all studies share a common true effect size, then *Q* would be distributed as chi-squared with degrees of freedom equal to the number of studies minus 1. So we can evaluate *Q* with reference to chi-squared to get a *p*-value. If *p* is less than alpha (typically set at 0.10 for this test) we reject the null, and conclude that some of the dispersion reflects variation in true effects.

Computing I^2

We can define I^2 as

$$I^2 = \frac{Q - df}{Q} .$$

In the numerator, since Q is the total sum of squares while df is the sum of squares attributed to sampling error, the difference is the sum of squares due to variance in true effects. In the denominator, Q is again the total. So, I^2 is the ratio of true to total.

Q and df are on a standardized scale. To convert either of these numbers to the metric of the effect size we would divide by C, a value based on the study weights. If we divide the numerator by C we get T^2, and if we divide the denominator by C we get V_{OBS}. So, we can rewrite the equation as

$$I^2 = \frac{T^2}{V_{OBS}},$$

which is the formula presented in the text. Equivalently, we could write this as

$$I^2 = \frac{T^2}{T^2 + V_{ERR}} = \frac{V_{TRUE}}{V_{TOTAL}},$$

which may be more intuitive.

Computing T^2

We can use Q and df to compute an estimate of the variance of true effects, T^2, using

$$T^2 = \frac{Q - df}{C}$$

In this formula the numerator is the sum of squares that reflects variation in true effects, but it is on a standardized scale. C is a factor based on the study weights that we applied when we standardized the deviations. Concretely,

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

where W_i is the weight for study i, which is $1/V_i$, the within-study error variance for that study. When we divide by C we reverse that process, so that T^2 is in the same metric that was employed for the synthesis.

The standard deviation of true effects, T, is then

$$T = \sqrt{T^2} \ .$$

Finally, the prediction interval is based on the mean plus or minus T.

Note that there are other ways to compute an estimate of the variance of true effects, T^2. The method described here was proposed by DerSimonian and Laird (DerSimonian & Laird, 1986).

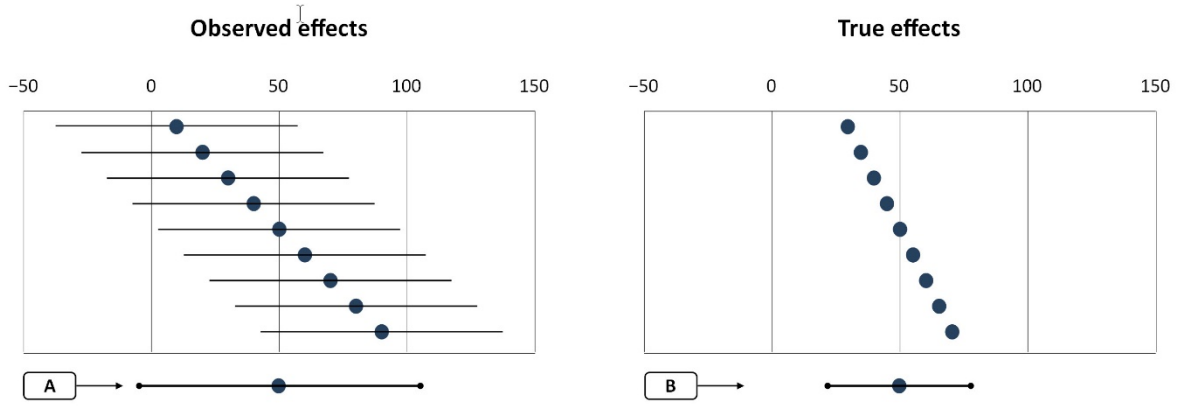


Figure 213 – Observed effects and true effects for a fictional meta-analysis

Figure 213 displays a fictional meta-analysis where the error variance is the same in all studies.

The left-hand plot shows the *observed* effects. This is the plot that is typically included with a published meta-analysis. The standard deviation of the observed effects is 27.4, and we expect that some 95% of all observed effects will fall within two standard deviations of the mean. This corresponds to a range of 110 points (−0.05 to +1.05) as suggested by Line [A]. The variance of the observed effects is the standard deviation squared, or 750.

By contrast, the right-hand plot shows the *true* effects for the same analysis. This is the plot that we *would* see if every study had an extremely large sample size, so that the sampling error was close to zero. This is the plot that we care about, since it tells us how much the effect size actually varies. The standard deviation of the true effects is 13.69, and we expect that some 95% of all true effects will fall within two standard deviations of the mean. This corresponds to a range of 55 points (23 to 77) as suggested by Line [B]. The variance of the true effects is the standard deviation squared, or 187.5.

Note. These schematics are intended to show that the true effects tend to fall closer to the mean than the observed effects. To make that point as clearly as possible we've drawn a uniform distribution, but in a real data set the effects would be expected to follow a normal distribution. We've also drawn the effects from smallest to largest. We do not mean to imply that the rows in the left-hand plot and the right-hand plot refer to the same studies.

In particular, we have some statistics that address the variance in observed effects, some that address the variance in true effects, and some that address the relationship between the two. This outline provides the context for the discussion that follows. In chapter ____ we discuss these statistics in detail.

For the present discussion, the key point is that the variance of the observed effects tends to be larger than the variance of true effects. To understand why that's true, consider what would happen if the true effect-size was identical in all studies. While the *true* effects would all be the same, the *observed* effects would vary because of sampling error. Concretely, the expected variance of observed effects (V_{OBS}) would be equal to the typical error variance (V_{ERR}). That is,

$$V_{OBS} = V_{ERR} . \tag{1.131}$$

The same paradigm applies when the true effects vary, as they do here. If T^2 is the variance of *true* effects, then the expected variance of the *observed* effects is given by

$$V_{OBS} = T^2 + V_{ERR} . \tag{1.132}$$

In words, the variance of observed effects is equal to the variance of true effects plus variance due to sampling error.

Statistics that quantify variation in the *observed* effects

The *Q*-statistic refers to the left-hand plot, and is defined as the sum of squared deviations (of each *observed* effect from the mean effect) on a standardized scale. If all studies share a common true effect size (and all the variance in observed effects is due to sampling error), *Q* would follow a chi-squared distribution with degrees of freedom equal to the number of studies minus 1. We can use this to obtain a *p*-value for a test of the null hypothesis that there is no variation in true effects. In this example *Q* is 10.67 with 8 degrees of freedom. The *p*-value for a test of the null (that all studies share a common true effect size) is 0.22.

Statistics that quantify variation in the *true* effects

The statistic called *T* (Tau) is the standard deviation of true effects. As such, it serves the same role as the standard deviation in a primary study. We can use the mean plus or minus two standard deviations to compute the prediction interval. If the effects are normally distributed, then the true effect size in some 95% of all comparable populations will fall within this interval. In this example, *T* is 13.69. The mean of 50 plus or minus two standard deviations yield a prediction interval of approximately 23 to 77 (line B in Figure 213)

The statistic called T^2 is the variance of true effects. This is simply the standard deviation squared. As is true in a primary study, the variance is not a terribly intuitive measure since it employs squared units rather than linear units. However, it has statistical properties that make it useful in the computations. In particular, the variance is a component in the weight assigned to each study for purposes of computing the mean effect size. In this example, T^2 is 187.50.

Statistics that quantify the relationship between the true and observed effects

If the right-hand plot shows the variance of true effects and the left-hand plot shows the variance of observed effects, it might be useful to have a statistic that quantifies the relationship between the two. This statistic is called I^2, defined as the ratio of true to total variance,

$$I^2 = \frac{V_{TRUE}}{V_{TOTAL}} = \frac{V_{TRUE}}{V_{TRUE} + V_{ERR}} = \frac{T^2}{V_{TOTAL}} \tag{1.133}$$

If I^2 is the ratio of true to total variance, then it is also the proportion of the observed variance that would remain if we could somehow remove the sampling error from the plot and (by definition) the ratio of the variance in the left-hand plot to the variance in the right-hand plot.

Consider the ratio of line [B] to line [A], where these lines are based on the standard deviations of the two plots. The ratio of the standard deviations (I) is

$$I = \frac{S_{True}}{S_{Obs}} = \frac{T}{S_{Obs}} = \frac{13.69}{27.49} = 50\% . \tag{1.134}$$

If instead, we computed the ratio of the variances, we would get the value of I^2, which is

$$I^2 = \frac{V_{True}}{V_{Obs}} = \frac{T^2}{V_{Obs}} = \frac{187.50}{750.00} = 25\% . \tag{1.135}$$

In sum,

In a primary study the standard deviation (S) tells us how the scores are distributed. The corresponding statistic in a meta-analysis is T, the standard deviation of true effects.

In a primary study we use regression to explain the variance, S2. In a meta-analysis we use regression to explain T2, the variance of true effects.

In a primary study we compute the sum of squared deviations, SS, as an interim step in computing the variance. In a meta-analysis we compute the sum of squared deviations on a standardized scale, Q, as an interim step in computing the variance.

In a meta-analysis, we may also use Q to test the null hypothesis that all studies share the same true effect size. This null hypothesis has no analog in the primary study.

In a meta-analysis we compute I2, which tells us what proportion of the variance in observed effects is due to variation in true effects rather than sampling error. This has no analog in a primary study, where we treat the observed effects as being identical to the true effects.

Note.

We used an example of a primary study with one level of sampling, where the observed scores are treated as being identical to the true scores. It's also possible to design a variant of this study where each student is tested multiple times, and we distinguish between the observed scores and the true score for any given student. In that case, the analysis would resemble the one we've described here for a meta-analysis.

Earlier, we explained that in a primary study we may treat each person's observed score as being identical to that person's true score. By contrast, in a meta-analysis, we distinguish between the observed effect size and the true effect size for each study. This becomes a key point in computing the variance.

In a primary study, we simply compute the variance of the observed effects. In a meta-analysis we can compute the variance of the observed effects, but that tends to be larger than the variance of true effects. To understand the reason, consider what would happen if we ran ten studies by drawing ten samples from the same population. Assume that the studies are identical in all respects. By definition, they all share the same true effect size (which is the effect size in the population). Yet, the observed effects will differ because of sampling error. And so, the variance of of the observed effects will be greater than zero.

If we assume, for the purpose of this example, that all the studies have precisely the same error variance, then the expected value of the observed variance will be equal to the error variance of each study. Indeed, that's the meaning of the error variance. While the fact that the variance of observed effects will be greater than the variance of true effects might be most obvious in the case where the variance of true effects is zero, it holds true also when the variance of true effects is not zero. Concretely, the variance of observed effects will be equal to the variance of true effects plus the variance due to sampling error.

We can use this fact, to estimate the variance of true effects. There are various ways to estimate the variance of true effects, but they all start with the method described here, called the DerSimonian and Laird method, or the method of moments. The method is to compute the variance of the observed effects, and then subtract the amount of variance we would expect to see based on sampling error. The difference is the variance of true effects.

WHAT DO WE MEAN BY A UNIVERSE OF POPULATIONS

Before we turn to the heterogeneity in a meta-analysis, we will review the idea of heterogeneity in a primary study.

Consider a primary study where draw a sample of students in a school, and record the scores on a math test. To compute the mean, we would sum the scores and divide by N. To compute the variance, we would compute the sum of squares (the squared deviations of each score from the mean), and divide this by the degrees of freedom. Then we'd take the square root of the variance, which gives us the standard deviation.

The mean and standard deviation enable us to describe the distribution. If we assume that the scores are normally distributed, then we expect that some 95% of all students in the population will score within two standard deviations of the mean. If the mean is 50 and the standard deviation is 20, then most students will score in the range of 10 to 90 (two standard deviations on either side of the mean).

The variance is less intuitive (since it's in a squared metric) but it has some statistical properties that make it useful for additional analyses. If we want to understand why the scores vary, we would probably employ analysis of variance or multiple regression. Both work with (and attempt to explain) the variance in scores.

Our goals in a meta-analysis are similar to those in this primary study– we want to know the mean and the standard deviation (which will enable us to describe the distribution of effects) and also the variance (which we will try to explain using procedures similar to those in the primary study). While the goals are the same, there are some important differences in the computations.

One difference has to do with the weight assigned to each study for the purpose of computing the mean and the other statistics. In the primary study, each student contributes the same amount of information as any other, and therefore all studies are given the same weight in the analysis. By contrast, in the meta-analysis, studies that yield a more precise estimate of the mean (typically the larger studies) are given more weight. This is an important distinction, but it's not central to the discussion at hand, and so we address it in a later chapter.

The second difference relates to the difference between a true score and an observed score. In the primary study described here, we define the student's "true" score as being whatever score the student received on the test. By contrast, in the meta-analysis, we make a distinction between the observed effect size for any study, and the true effect size for that study. The *observed* effect size is the one that we see (the effect size that is recorded and plotted in Figure 1). By contrast, the *true* effect size is the one that we would see if we had somehow enrolled the entire population for a study. The observed effect size serves as an estimate of the true effect size but, because of sampling error, invariably falls below or above the true effect size.

Because of this, when we are working with the effect sizes in a meta-analysis, we are working with two distinct distributions of effect sizes. One is the distribution of observed scores, and the other is the distribution of true scores. And they are not the same.

When our goal is to estimate the mean effect size, the distinction between the two distributions is not terribly important. We compute the mean of the observed effects, and this gives us the expected mean of the true effects. This works, because the sample mean is an unbiased estimate of the true mean – it will be too low in half the cases, and too high in the other half.

By contrast, when our goal is to estimate the variance (or standard deviation) of the effect sizes, the distinction between the two distributions is of major importance. The variance of observed effects tends to be larger than the variance of true effects. Therefore, we cannot simply compute the variance of observed effects and use this in the analysis. Rather, we compute the variance of observed effects, then impute the variance of true effects, and use the latter in the analysis.

The reason that the variance of observed effects tends to be larger than the variance of true effects, and the methods we use to impute the latter, are explored in another chapter. For the present purposes, the key point is that they *do* differ, and this provides the background that we need for discussing the heterogeneity statistics. We have some statistics that reflect the variance of observed effects, some that reflect the variance of true effects, and some that reflect the relationship between two.

As it happens, the ratio of the two I2 will often be close to the ratio of the two T2. However

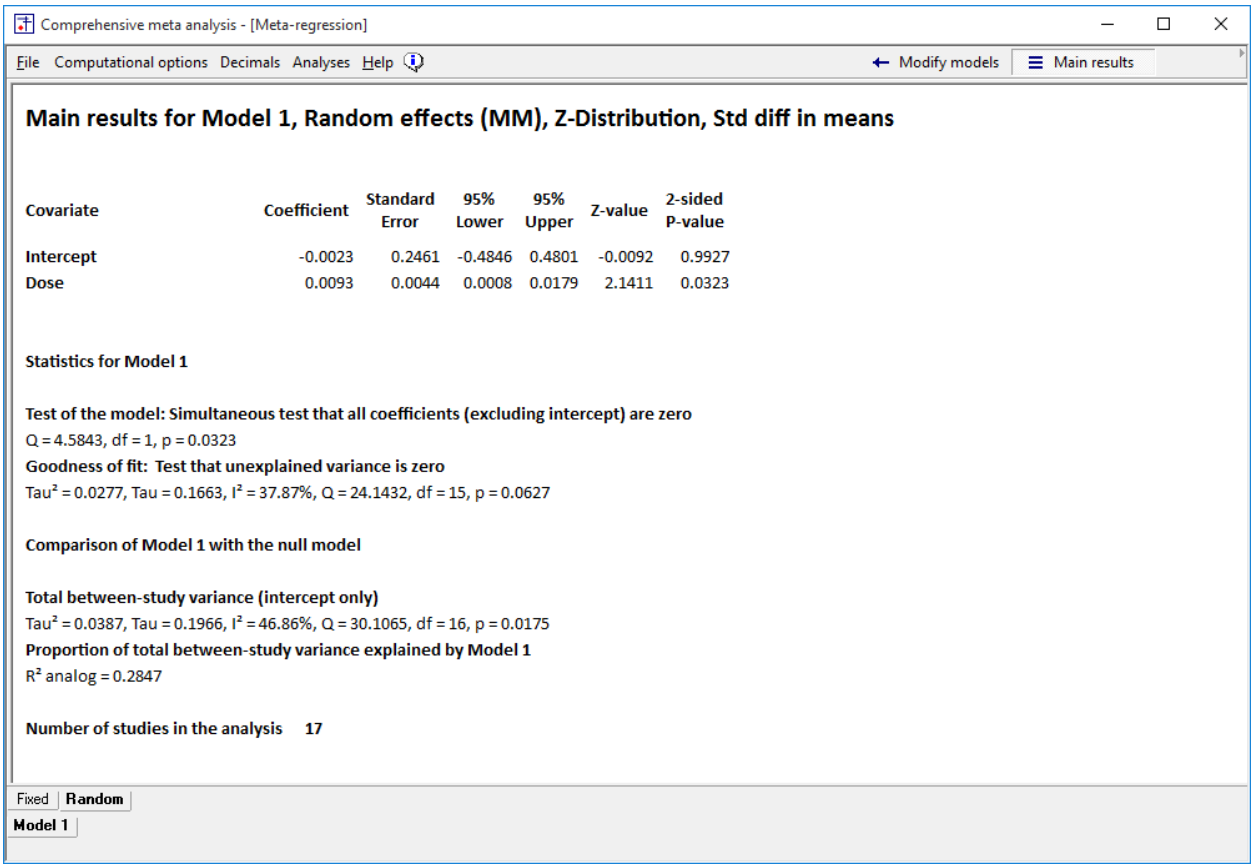
the variances. But if we understand what these numbers represent, it's clear we should not be using them to report how much variance is explained by the covariates. Rather, if we want to report how much variance is explained by the covariates we should work directly with the variances as discussed below.

In this example I2 was initially around 47% (section D) and then dropped to around 2% when we added covariates (Section C). It's common for I2 to drop as we add covariates, and the reason can be found in

$$I^2 = \frac{V_{TRUE}}{V_{TRUE} + V_{ERR}} \tag{1.136}$$

The variance of true effects about the regression line [C] should be smaller than the variance of true effects about the grand mean [D]. Indeed, the goal of the regression is to reduce the unexplained variance. At the same time, the within-study error variance should be the same in [C] and [D]. So, as we reduce VTRUE, the numerator decreases at a faster rate than the denominator, and I^2 drops.

While I^2 should always decrease as covariates are added, there will be some cases when it does not. This is because the variance of true effects is being estimated with error. If we underestimate T2 in section [D] and then overestimate T2 in section [C], it will appear that the value of T2 has increased, and the estimated value of I^2 will increase as well.



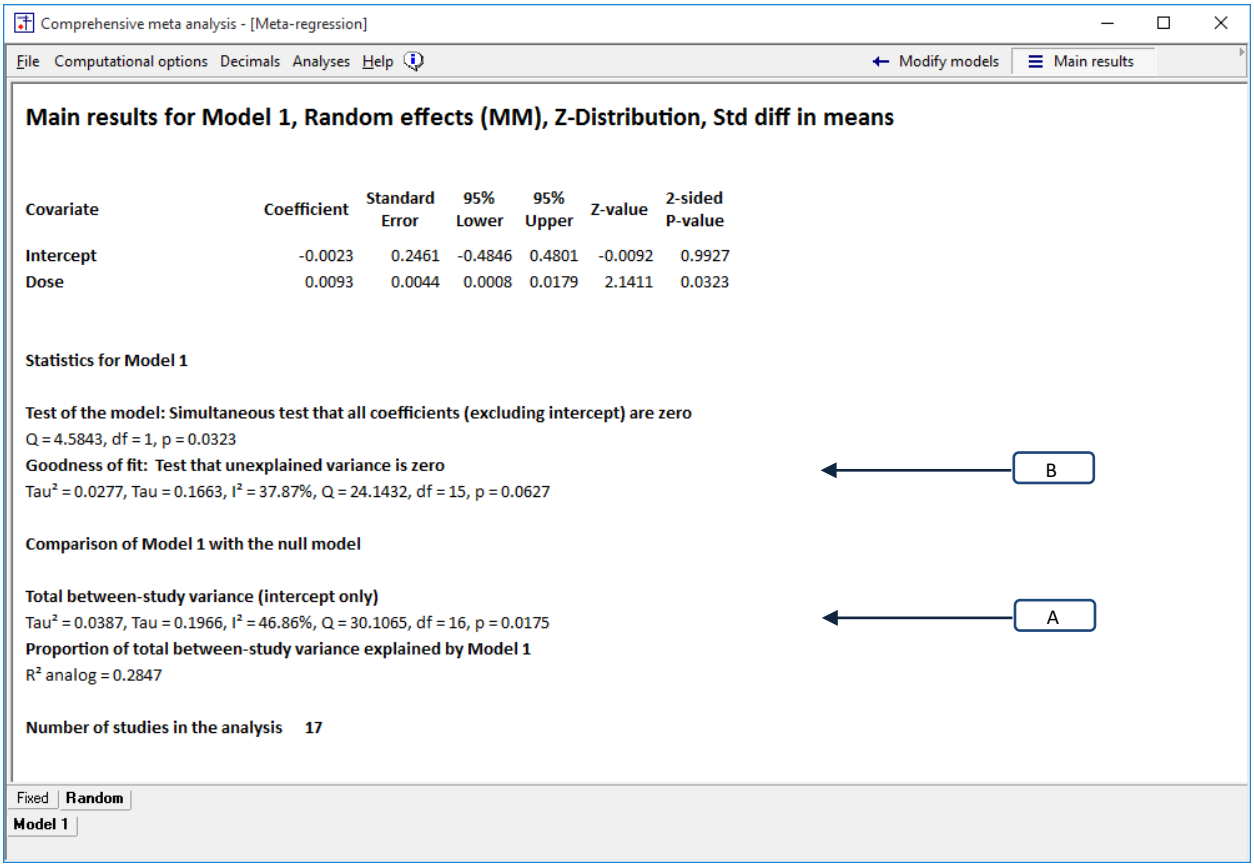


Figure 214 | Main results | Random-effects

Section [A] reports statistics for a regression line with no covariates. Here, T^2 is 0.0387 and I^2 is 46.86%. If we look at the dispersion of effects about this regression line (which is simply the grand mean), 46.86% of this reflects variation in true effects. The actual amount of variation in true effects is 0.0387.

Section [B] reports statistics for a regression line with Dose as a covariate. Here, T^2 is 0.0277 and I^2 is 37.87%. If we look at the dispersion of effects about this regression line, 37.87% of this reflects variation in true effects. The actual amount of variation in true effects is 0.0277.

Without a covariate [A] I^2 is estimated as

$$I^2 = \frac{T^2}{T^2 + V_{ERR}} = \frac{0.0387}{0.0387 + 0.0439} = 46.86\% \qquad (1.137)$$

With a covariate [B] I^2 is estimated as

$$I^2 = \frac{T^2}{T^2 + V_{ERR}} = \frac{0.0277}{0.0277 + 0.0454} = 37.87\% \quad (1.138)$$

Note that the difference between [A] and [B] is that we've added a covariate to the prediction equation, and that when we do this there is a decrease in both T^2 and I^2. The reason that T^2 drops is that the purpose of the covariate is to reduce the unexplained variance in true effects (which is the definition of T^2). Note that I^2 drops in tandem with T^2, because I^2 is defined as

$$I^2 = \frac{T^2}{T^2 + V_{ERR}} \quad (1.139)$$

and as T^2 drops, the numerator will fall more quickly than the denominator. In the extreme, if *T2* is zero then I^2 will also be zero.