CAMARADES Monograph No. 1

Combining the evidence from different drug trials

By

Qiuxun Wu

Dissertation Presented for the Degree of
MSc in Operational Research

2006

Abstract

In this project, a program of simulation is developed to investigate the power of the stratified meta-analysis, which is used to combine the evidence from different drug trials. The particular focus is the investigation of drugs for stroke treatment in animal trials, which usually have small sample sizes. Two approaches, the normalised difference in means and the standardised difference in means are used in the heterogeneity tests and compared under various circumstances.

Acknowledgements

First of all I would like to express my most deep and sincere gratitude to my supervisor, Dr Malcolm R Macleod, whose useful suggestion helps me all the time during the research and writing-up period. And I also want to appreciate another supervisor of mine, Prof. Colin G.G. Aitken for his help and guidance on the relative theoretical backgrounds.

I am grateful to another friend Mu Hu for her continual encouragement and concern to me during the whole project period. It is very lucky for me to meet her in Edinburgh.

Finally I owe my loving thanks to my parents who provide the chance for me to study abroad and offer me a great deal of love and care in the past 23 years.

Table of Content

TABLE OF CONTENT **4**

1. INTRODUCTION & BACKGROUND **5**

1.1. INTRODUCTION 5
1.2. BACKGROUND OF THE STROKE DRUG TRIALS 6

2. META-ANALYSIS **7**

2.1. GENERAL PRINCIPLE 7
2.2. ESTIMATES OF EFFECT SIZE IN A SINGLE STUDY 7
2.3. POOLED ESTIMATES OF EFFECT SIZE IN META-ANALYSIS 10
2.4. STRATIFIED META-ANALYSIS 12

3. THE SIMULATION OF THE STROKE DRUG TRIALS **14**

3.1. ASSUMPTIONS OF THE SIMULATION MODULE 14
3.2. PROCESS OF THE SIMULATION 15
3.3. LIMITATIONS TO THE SIMULATION 18

4. THE PROGRAM OF THE SIMULATION **19**

4.1. THE GENERAL INTRODUCTION TO THE PROGRAM OF THE SIMULATION MODULE 19
4.2. THE STRUCTURE AND CONTENT OF THE PROGRAM 20

5. THE BEHAVIOR OF THE SIMULATION MODULE **23**

5.1. THE ESTIMATES OF EFFICACY UNDER NMD & SMD METHOD 23
5.2. POWER TO DETECT HETEROGENEITY BETWEEN GROUPS 24
5.3. SENSITIVITY ANALYSES OF THE HETEROGENEITY INVESTIGATION 26
5.3.1. Sensitivity to effect size *27*
5.3.2. Sensitivity to the base effect size *28*
5.3.3. Sensitivity to the significant level *30*
5.3.4. Sensitivity to the individual numbers of a single study *31*
5.3.5. Sensitivity to the subgroups' sizes *33*
5.3.6. Sensitivity to the evenness of the subgroups' sizes *34*

6. CONCLUSION & DISCUSSION **37**

6.1. CONCLUSION 37
6.2. DISCUSSION 38

REFERENCES **40**

Appendices……**38**

1. Introduction & Background

1.1. Introduction

When a new drug is being investigated, animal experiments are done before the clinical trials and hundreds of data of the efficacy are obtained. These animal data are the results from many different drug trials, which all use the same drug but at different doses, given at different times, taken by different species, represented with different terms of measurement of efficacy and done in different labs. All these differences make it difficult to tell how effective the new drug is at first glance. Hence, it is necessary to have a full and unbiased assessment based on all available animal data to show an overview of the efficacy of the drug itself as well as the limits to that efficacy.

Meta-analysis is used to combine the evidence from different studies to get an overall estimate of the drug efficacy. However, we do not combine the data from all of the trials as if they were from a single large trial. Such an approach is inappropriate for several reasons and can give misleading results, especially when the number of participants in each group is not balanced within trials. [1] The pooled estimate result will lose all the information on study characteristics from each individual trial, characteristics which might have great influence on the efficacy of the drug. Such a global efficacy does not provide sufficient information to allow the assessment of the new drug on which to base the decision to proceed to the clinical trials. To achieve this, we can use the stratified meta-analysis to give a more detailed assessment than a global estimate. In this case, the heterogeneity investigation is also necessary and important to carry out.

The particular focus for this project is the development of drugs for stroke treatment in humans. A simulation is carried out by using stratified meta-analysis. Two different measurements of effect size, the normalised mean difference and standardised mean difference, are compared to find out which one performs better under various circumstances, especially when -as is typical of animal experiment-the sample size is small. A program of the simulation was developed so that the process of the simulation can be repeated for thousands and hundreds of times to get the power of the methods used. In addition, several sensitivity analyses were also be done to investigate the

limitation of the meta-analysis.

1.2. Background of the stroke drug trials[2]

The specific area this project concerns about is the candidate drugs for stroke treatment in humans. The object of these trials is to find out how effective a drug is at reducing brain damage after the blood supply to part of the brain has been cut off. Typically, several animals of same species were divided into two groups: the treatment group and the control group. In the former group, the animals were treated with the drug before or/and after ischaemia, while the animals in the other group were not treated. There were two terms of outcomes of the trials. One was measured as a neurologic score which get from the neurological test; the other was measured as infarct size which can be assessed by measuring the vitality unstained area in the slice of brain as an area of death and integrating this to give a volume of infarction with slice thickness. This whole process of the work was repeated by various different research groups with same drug but with different animals under different circumstances.

To get the effect size of the drug, a ‘comparison’ is undertaken with the outcomes from the trials. The ‘comparison’ is defined as the assessment of outcome in treatment and control groups after the treatment of the stroke. For each comparison, data are extracted for mean outcome, standard deviation and number of animals per group. Then, such data are aggregated using random effects model of DerSimonian and Laird(DSL). Stratified Meta-Analysis is also used to explore the impact of various characteristics on estimating effect size. Then, the overall estimate of the effect size derived from the DSL model is reported as the efficacy of the drug on animals along with the detailed limits to it obtaining from the stratified meta-analysis.

2. Meta-Analysis

2.1. General principle

When a series of studies are expected to share a common effect size, we can consider combining all the studies to obtain a pooled estimate of the effect size. The method that we usually use is called Meta-Analysis. Meta-Analysis is a process of two stages. In the first stage, a summary statistic of each study is calculated, which can be risk ratios, odds ratios or risk differences for event data, difference in means for continuous data, or hazard ratios for survival time data.[3] In the second stage, all the effect sizes, the summary statistic of each study gained in the first stage, are pooled to get a overall statistic to describe the common efficacy. There are two methods to obtain the combination of estimates of effect size[4]: (i) a direct weighted average of the summary statistics and (ii) a maximum likelihood estimator. When concerning the weights in the method (i), they must to reflect the amount of information contained in each single study. In addition, the confidence interval and statistical significance of the pooled estimate are also calculated. All these mentioned above are the general principles that commonly used methods of Meta-Analysis follow.

In this project, we are dealing with continuous data; therefore the summary statistics we calculated in the first stage are the difference in means. In the second stage, since the two methods are equivalent and the former one is simpler and involves less computation, it is used more frequently than the latter one. In this case, the second method, the maximum likelihood estimator will not be discussed here.

2.2. Estimates of effect size in a single study

In a single study, before we calculate the summary statistics, several other statistics are required: the number of participants, the mean of the outcomes and its standard deviation for both control and treatment groups. (See table 2.1)

Study i	**Group Size**	**Mean**	**Standard Deviation**
Treatment	$\mathbf{n_{ti}}$	$\mathbf{m_{ti}}$	SD_{ti}
Control	n_{ci}	m_{ci}	SD_{ci}

Table 2.1 information required for calculating summary statistics

And the pooled standard[3] deviation of the two groups is given by

$$s_i = \sqrt{\frac{(n_{ti}-1)SD_{ti}^2 + (n_{ci}-1)SD_{ci}^2}{N_i - 2}},$$

where N_i is the sum of n_{ti} and n_{ci}.

There are many ways to calculate the summary statistics of continuous outcomes. For instance,

(i) Simple difference in means (denoted MD)[5]

$$MD_i = m_{ti} - m_{ci},$$

with standard error

$$SE(MD_i) = \sqrt{\frac{SD_{ti}^2}{n_{ti}} + \frac{SD_{ci}^2}{n_{ci}}}$$

It can be used when all the outcomes of the trials are measured on the same scale.

(ii) Normalised difference in means (denoted NMD)

$$NMD = \frac{m_{ti} - m_{ci}}{m_{ci}} \times 100,$$

with standard error

$$SE(NMD_i) = \sqrt{\frac{SD'^2_{ti}}{n_{ti}} + \frac{SD'^2_{ci}}{n_{ci}}}, \text{ where } SD'_{ti} = \frac{SD_{ti}}{m_{ci}} \times 100 \text{ and } SD'_{ci} = \frac{SD_{ci}}{m_{ci}} \times 100.$$

Actually, the normalised difference in means is the modified (or advanced) form of the simple difference in means. The original outcomes are processed to the ratio of the mean of control group, so that the same outcomes measured in different scales can also be combined.

(iii) Standardised difference in means (denoted SMD)[6]

There are various formulae of effect size used in the standardised mean difference method. They differ with respect of the standard deviations used in the formulae or if a correction of small sample bias is included. The formulae showed below are two popular ones.

Cohen's d[6] is given by

$$d_i = \frac{m_{ti} - m_{ci}}{s_i},$$

with standard error

$$SE(d_i) = \sqrt{\frac{N_i}{n_{ti} n_{ci}} + \frac{d_i^2}{2(N_i - 2)}}.$$

Hedges' adjusted g[6] is defined as

$$g_i = \frac{m_{ti} - m_{ci}}{s_i}\left(1 - \frac{3}{4N_i - 9}\right),$$

with standard error

$$SE(g_i) = \sqrt{\frac{N_i}{n_{ti} n_{ci}} + \frac{g_i^2}{2(N_i - 3.94)}}.$$

The Hedges' adjusted g is very similar to Cohen's d and the only difference is the adjustment added to correct the small sample bias.

In this project, we use the latter two methods, the normalised mean difference and the standardised mean difference, to calculate the summary statistics. For the standardised mean difference, the Hedges' adjusted g is used because the number of animals per group is usually fewer than 10.

The advantage of standardised mean difference is that it can combine outcomes measuring the same underlying effect and ignore differences in the scales used. This method is validated to be powerful and promising in the clinical trials, in which case the number per group is usually very high (200 to 1000) so that the observed standard deviation used in the formulae is quite close to the population standard deviation. However, in the animal drug test, the number of animals per group is much smaller (usually 5 to 10), and the observed standard deviation cannot be regarded as a precise estimate of the population standard deviation. Under this circumstance, it might lead to an overestimate of the effect size.[3]

2.3. Pooled estimates of effect size in Meta-Analysis

Generally, the summary statistics of each single study are denoted by θ_i with a weight w_i .According to different summary statistics obtained in the individual study, various methods of calculating the overall estimate of the effect size can be used correspondingly. All these methods can be classified into two categories, fixed effect and random effects methods, depending on the assumption made of the true effect size. In the fixed effect Meta-Analysis it is assumed that the true effect of the treatment is the same value in each individual study, or fixed, the differences between study results being due solely to the play of chance.[3] In the random effects Meta-Analysis the treatment effects for the individual studies are assumed to vary around some overall average treatment effect.[3] Since the effect size in each single study we gained above is mean difference, only two approaches used in the simulation are described here.

(i) Inverse variance method

Inverse variance method is a fixed effect method and can be used to pool either binary or continuous data and therefore is widely applied. The point estimate of the overall efficacy is given by

$$\theta_{IV} = \frac{\sum w_i \theta_i}{\sum w_i} .$$

The weights are decided by the square of the standard errors:

$$w_i = \frac{1}{SE(\theta_i)^2} .$$

It is not difficult to note that the standard error of the summary statistics depends on the size of the sample and the quality of the research. The outcome obtained from high quality research and large sample which means small standard error should give more weights so that the combined estimate of the efficacy can be more accurate or close to the true value. The choice of the weight in this approach minimizes the variability of the pooled point estimate of the effect size.[3]

The standard error of the point estimate θ_{IV} is given by

$$SE(\theta_{IV}) = \frac{1}{\sqrt{\sum w_i}}.$$

The heterogeneity statistic is given by

$$Q = \sum w_i(\theta_i - \theta_{IV})^2 .$$

(ii) DerSimonian and Laird random effects models

Under the DerSimonian and Laird random effects model, the effect sizes θ_i derived from each individual study are assumed to be normally distributed with mean and variance τ^2. The usual DerSimonian and Laird[7] estimate of τ^2 is given by

$$\tau^2 = \frac{Q - (k-1)}{\sum w_i - \left(\frac{\sum w_i^2}{\sum w_i}\right)},$$

where Q is the heterogeneity statistic calculated in the inverse variance estimate, k is the number of individual effect size combined in the Meta-analysis. If Q < k-1, which indicates that the heterogeneity is smaller than its degrees of freedom, τ^2 is set to zero. The weight of each study's effect size is given by

$$w'_i = \frac{1}{SE(\theta_i)^2 + \tau^2}.$$

Compared to the inverse variance method, the adjustment factor τ^2 is added when choosing the weight.

Thus the overall effect size is given by

$$\theta_{DL} = \frac{\sum w'_i \theta_i}{\sum w'_i},$$

with standard error

$$SE(\theta_{DL}) = \frac{1}{\sqrt{\sum w'_i}}.$$

Note that when τ^2 is zero, which means the heterogeneity among the studies combined is very small, the weights are the same as those given by inverse variance method. When the

estimate of τ^2 is greater than zero, the weights in the random effects models will be smaller and more similar to each other than those calculated in the fixed effect models. It also can be noticed that the random effect models give relatively more weight to smaller studies than the fixed effect models.

The formulae used in the two models described above are the modified version in practice use. Note that the weights are given by the inverse of the standard error of the effect size instead of the variances required in the general form. In this case, the estimator of the overall effect size is no longer unbiased. Brockwell and Gordon[8] write

"For both the fixed and random effects methods, inference is carried out ignoring the sampling errors in the individual study variances. Estimated values $\hat{\sigma}_i^2$ are used without modification to the form of $\hat{\mu}$, its variance or distribution." ($\hat{\sigma}_i^2$ are the estimated values of true variance and $\hat{\mu}$ is the overall estimate θ mentioned above.) This issue is specially a problem for small sample sizes in which case the estimates of the variances have low precision.[9] However, the problem diminishes if sample sizes increase.[9] The issue just mentioned is really a problem to the animal drug test because of the small numbers of animals used.

2.4. Stratified Meta-Analysis

Sometimes it is not proper to combine all the single studies to obtain an overall estimate of the efficacy since the process of the combination will conceal the features or characteristics which might have a great influence on the effect size. Combining the outcomes regardless of the issue will probably lead to an overestimate of the efficacy. Under such circumstance, stratified meta-analysis can be used to investigate the heterogeneity of the data and its impact on the overall estimate in order to give a full and unbiased assessment of the efficacy of the drugs. In the stratified meta-analysis, the studies are grouped into a small number of categories according to a particular feature or characteristic, and then a separate meta-analysis is carried out within each subgroup.

An inference that the treatment effect differs between two or more subsets of the trials should be based on a formal test of statistical significance.[3]

If there are only two subgroups, a z-test can be carried out to test the significance of the difference between them. In the test, we compare the z statistic[3]

$$z = \frac{\theta_1 - \theta_2}{\sqrt{[SE(\theta_1)]^2 + [SE(\theta_2)]^2}},$$

with critical values of the Normal distribution.

An alternative test is the Chi-square test which can be used regardless of the number of the subgroups. The idea is that the overall un-stratified heterogeneity is partitioned into the part explained by differences between subgroups and the part remaining unexplained within the subgroups. Let Q_T, Q_B, Q_k denote the overall un-stratified heterogeneity, the heterogeneity between groups and the heterogeneity of the *k*th group respectively. The Q_B is given by[3]:

$$Q_B = Q_T - \sum_k Q_k \ .$$

Then, compare this Q_B with critical values of the chi-squared distribution with k-1 degrees of freedom.

The heterogeneity tests should be treated with caution for some reasons. Observing the formula used to calculate the heterogeneity, it is obvious that the investigation of differences between subgroups is non-randomized no matter whether the fixed effect method or the random effects method is used. And the significant differences can easily arise by chance or other factor. When multiple possible sources of heterogeneity are investigated, the chance that one of them being found to be statistically significant increases. Hence, the number of factors considered should be restricted.[3]

The stratified meta-analysis used in the stroke drug trials is pre-specified. It increases the credibility of statistically significant findings but might ignore a particular characteristic that does influence the efficacy of the drug. One possible solution is to use the pre-specified meta-analysis first and then consider other features which might impact the effect size.

3. The simulation of the stroke drug trials

In order to demonstrate whether or not the small sample size is a problem in practice, and which method, the normalised mean difference meta-analysis (NMD) or the standardised mean difference meta-analysis (SMD), performs better under various circumstances, a simulation is carried out to get a straight illustration. The main purpose of the simulation is to test and compare the power of the NMD and SMD by using the stratified meta-analysis.

In addition, the simulation can also give a reference to the people who are working on a investigation of one drug. They can use the simulation to get the power of the research with the parameters they used in the research. For example, a research group tested a drug on 10 animals which were divided into two groups evenly. The mean and standard deviation is also known. They can input these data into the simulation and to see if the power is high enough to say the research result is promising or more tests are needed to increase the power.

3.1. Assumptions of the simulation module

Before building the simulation module, there are a few of assumptions should be made.

(i) In each subgroup, the outcome from each individual animal in a single study is normally distributed, thus the effect size θ_i is also normally distributed. Since the DerSimonian and Laird random effects models are used to combine the data in the subgroups, the simulation have to follow the assumption the model required.

(ii) The outcomes of each group are independent to each other. The data in the treatment or control group are not related, which means they are generated from independent random numbers.

(iii) All the data are grouped into two subgroups according to whether a certain characteristic, say Y is involved. And the true value of the effect size in the subgroup with Y is always large than the one in the subgroup without Y. Actually, this assumption are not restrictedly required but just made for the sake of simplicity.

(iv) The means of the outcomes in both control groups are the same regardless of the presence of the feature Y. It is reasonable because the animals in the control groups will not get any treatment (no drug will be given), which means no influence on the efficacy of the drug.

(v) The numbers of animals per group and the population's standard deviations are the same in either control groups or treatment groups, but vary between control and treatment groups. Once more, this assumption is made for the sake of simplicity and can be relaxed by a slight modification in the simulation program.

3.2. Process of the simulation

The whole process of the simulation is described as following.

1. Decide respectively the number of animals per group for the control or the treatment group, say n_t and n_c.
2. Generate n_t and n_c random numbers between 0 and 1(uniform distributed) for both subgroups with and without characteristics Y.
3. Calculate the inverse of the standard normal cumulative distribution of the random numbers from the previous step.
4. For the control groups of both subgroups with and without characteristics Y, multiply the inversed random numbers by the population standard deviation and then add the mean of this group. The population standard deviations and the means are the same.
5. Repeat the same calculation to the inversed random numbers in the treatment groups of both subgroups but with different values of means and population standard deviations.
6. Calculate the observed average, standard deviation and pooled standard deviation in all groups.
7. Do the comparison between the control and treatment groups. Calculate the values using normalised difference mean as well as the standardised difference mean methods.
8. Repeat the steps 1-7 a certain times to get a sample of the subgroup without characteristics Y.
9. Repeat the steps 1-7 another certain times (can be the same as the times in the previous step) to get a sample of the subgroup with characteristics Y.
10. Combine the data using the stratified meta-analysis to obtain the overall estimates of effect size for each subgroup and the whole data.
11. Calculate the total heterogeneity and the heterogeneity within each subgroup. Then, subtract the

$Q_{within\ group}$ from the Q_{total} to get the heterogeneity between the subgroups.

12. Compare the observed heterogeneity between subgroups with the critical value of chi-squared distribution with 1 degree of freedom to find out if it is a significant difference between the two subgroups.
13. Use the DerSimonian and Laird random effects method to calculate the estimates of effect size in the two subgroups. And then perform the z-test to investigate the difference between the groups.
14. Repeat the whole process mentioned above thousands of times.
15. Work out the proportion of times the test told the difference, i.e. the power, using t-test and chi-squared test.

For the simulation of a single study in the first stage of the meta-analysis:

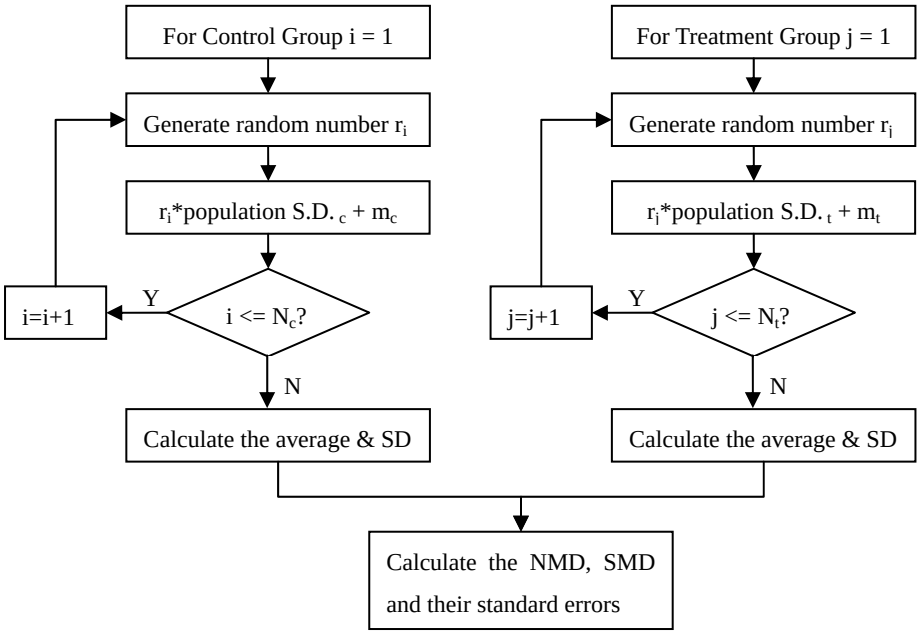


Chart 3.1:m_c, m_t, are the population means of the control and treatment group respectively. N_c, N_t are the numbers of animals per group.

For the simulation of combination in the second stage of the meta-analysis:

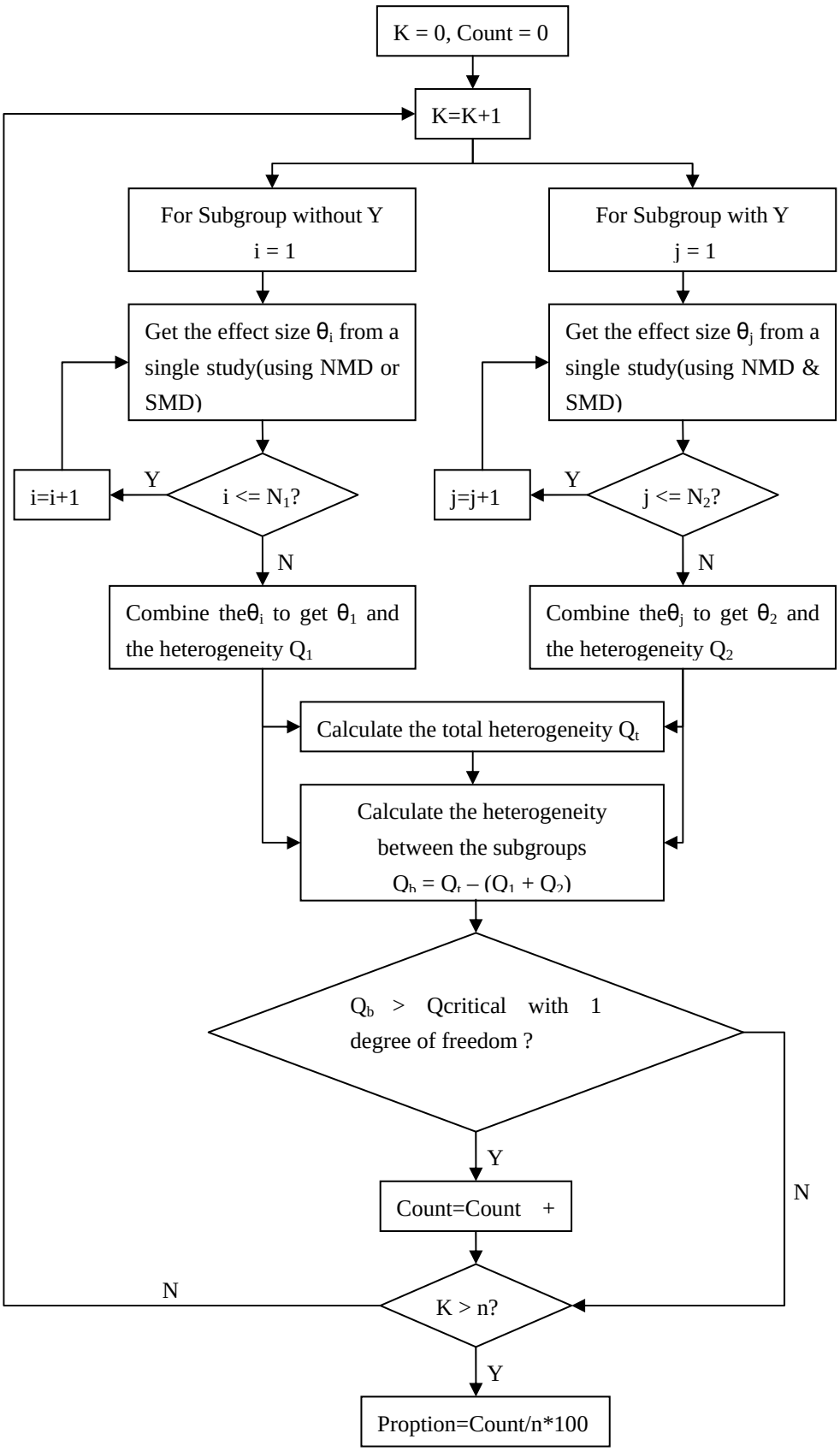


Chart 3.2: N_1, N_2 are the sample size of the two subgroups respectively. n is the times of run of the simulation required by the user.

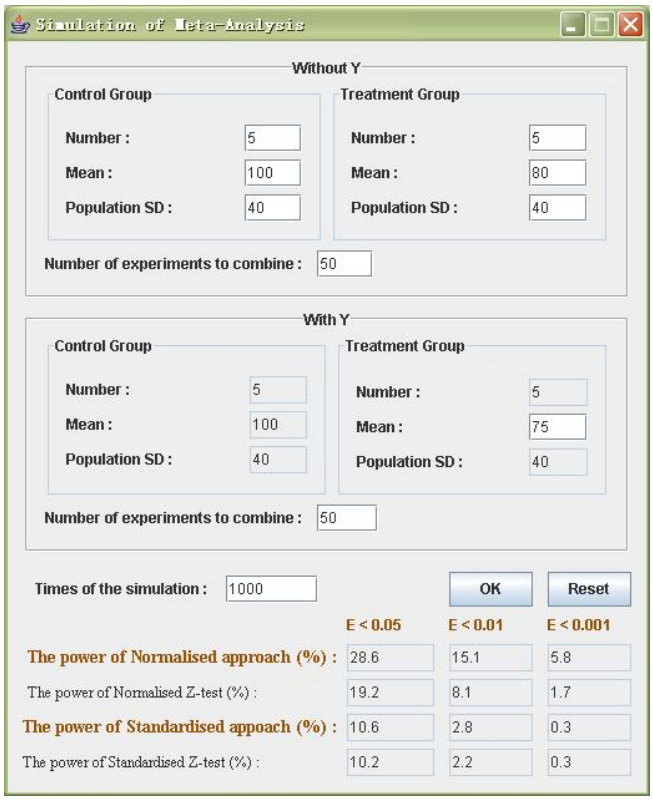
3.3. Limitations to the simulation

Owing to the nature of the computer program, there are some limitations to the simulation. Firstly, the random numbers generated by the computer are not truly random but rather pseudorandom and are not completely independent to each other. In some cases, a bad seed will lead to a series of numbers with high correlations. Secondly, the algorithm used to calculate the inverse of the standard normal cumulative distribution has a small estimate error from the true value as the computer can only handle discrete data. The way it calculate the integration makes it impossible to avoid such errors. Thirdly, the detailed outcomes of the simulation are not available unless the source code is modified. (In fact, the action of output the details will slow down the program a lot which is not desirable and the output of the details is also not required.) In some cases, the powers of the NMD and SMD are not enough to understand what is going on underlying and the simulation seems to be lack of evidences. Finally, the speed of the simulation program is not as fast as expected when the sample size is large. It might have no response if a huge number is input as the number of animals per group, times of experiments or the times of simulation.

4. The program of the simulation

4.1. The general introduction to the program of the simulation module

I developed a computer program coded in Java language to realize the simulation. It is named as Meta-Analysis_simulation.jar and a relatively simpler version is also available as an applet. Another two libraries, which are javaws.jar and swing-layout-1.0.1.jar, are required when running the program. The former one is used to contain the information and structure of the whole program while the latter is used to support the display of the Graphic User Interface (usually known as GUI). When using the program, the Meta-Analysis_simulation.jar and the two libraries should be regarded as a whole self-contained program in the folder named 'dist' (available in the CD submitted), and the relative directories position in it cannot be changed otherwise it will report errors and the program will not work. The figure 4.1 is a screenshot of the program.



The values of the 'numbers' and the 'times of the simulation' should be integer. Other parameters should be positive real digits. The gray text fields are not editable but will change when the corresponding white text fields are edited. To run the simulation, just input the parameters and press the 'OK' button. The 'Reset' button is used to set all the values of parameters to the default ones.

figure 4.1 screenshot of the program of simulation

NetBeans 5.0 was used as an IDE (integrated development environment) to develop the program. It is an open-source software available free of charge and is available for downloading on *http://www.netbeans.org*. The development of the GUI was much simpler as part of its source code

was generated by the NetBeans automatically, which was the blue part in the source code when it was opened in the NetBeans.

4.2. The structure and content of the program

The program consists of four classes, which are ‘Simulation_Of_MetaAnalysis.java’ (‘simulationJApplet.java’ for the applet), ‘StatUtil.java’, ‘Trials.java’ and ‘MetaAnalysis.java’.

‘StatUtil.java’ is the class used to calculate the inversed of the standard normal cumulative distribution. The original source code was found through the internet. It was developed by Sherali Karimov (*sherali.karimov@proxima-tech.com*) basing on the algorithm written by Peter J. Acklam (*jacklam@math.uio.no*). The class contains the implementation of:

* - Inverse Normal Cumulative Distribution Function Algorithm
* - Error Function Algorithm
* - Complimentary Error Function Algorithm

Only the first implementation was used in my program.

‘Trials.java’ is the class used to generate and process the data of a single study of animal drug test. The single study is regarded as an object with the properties of two groups, the control group and the treatment group. It contains the implementation of:

* - Construction of the class and Generator of the single study

  The required numbers of random numbers are generated by using the *Math.random* and then processed as the description in the step 2-5 in chapter 3.2 to form the sample of a single trial.

* - Calculate the averages, standard deviations of the control and treatment group
* - Calculate the Normalised effect size and its standard error
* - Calculate the Standardised effect size (using Hedges' adjusted g) and its standard error
* - Get the values of control and treatment group

  It allowed the external classes or functions to access the value of control and treatment group.

‘MetaAnalysis.java’ is the class used to do the combination in meta-analysis. The process of the second stage in meta-analysis (combination part) is regarded as an object. The class contains the

implementation of:

* - Construction of the class and Generator of the subgroup

  Several functions of the 'Trials.java' are recalled to generate the effect sizes and the corresponding fixed weights which will be combined in the stratified meta-analysis of a subgroup.

* - Calculate the random effect size of one subgroup

* - Combine the data in one subgroup and calculate its overall effect size.

* - Calculate the 95% confidence interval for the overall effect size

* - Get the values of the effect sizes and the weights

  It allowed the external classes and functions to access the value of the effect sizes and their weights (Both fixed and random ones are available.)

'Simulation_Of_MetaAnalysis.java' is the main class of the whole program to run the simulation though it is actually a GUI (Graphic User Interface) that allows users to input parameters in a windows frame. The parameters needed in the simulation are input by the users and transfer into the program through the GUI. Then the process described in chapter 3.2 is operated. The power of NMD and SMD method is tested with Z-test as well as the chi-square test at the confidence levels of 5%, 1% and 0.1%. The results are displayed in the GUI frame. There is only one function in the class: 'Heterogeneity (double [] weight, double [] theta)', which is used to calculate the heterogeneity of a group of effect sizes.

Note that not every function written is used in the main class of the simulation program, especially the ones used to print out the details of the simulation (the part commended in the source code). Because it is not necessary to view all these data and printing out all of them will definitely slow down the speed of the simulation which is not desirable. If there is any need for the details, just uncomment the part which is used to print out the details in the source code. Then, each time the 'OK' button is pressed, the details will be output into an excel file called 'output.xls'. However, one more library named 'jexcel' is required and only the values generated in the last time of simulation are output and it takes a longer time. Hence, it is better to set the times of the run of the simulation as once to save time.

5. The behavior of the simulation module

After running the simulation hundreds of and thousands of times, a general picture of the behavior of the simulation module was drawn. The whole chapter gives the idea how the NMD & SMD meta-analysis perform under various circumstances when a stroke drug is testing on animals.

5.1. The estimates of efficacy under NMD & SMD method

As mentioned in the assumptions in chapter 3.1, the true values of the efficacy are supposed to follow the normal distribution with the mean μ and the standard deviation σ. In the program of the simulation, the outcomes of either control group or treatment group are generated normally distributed with uniform random numbers. Thus, the effect size of each single group is also normally distributed. The simulation was run for several times to get a general idea of what the effect sizes are when the individual number of animals per group is small.

The charts showing below are the frequency of the results coming out of a simulation run with 5 individual animals in both control and treatment group. The mean and population standard deviation of the control group are 100 and 40 respectively, while those of the treatment group are 80 and 40. There are 500 studies were generated and the effect sizes were calculated with Normalised Difference in Means as well as Standardised Difference in Means.

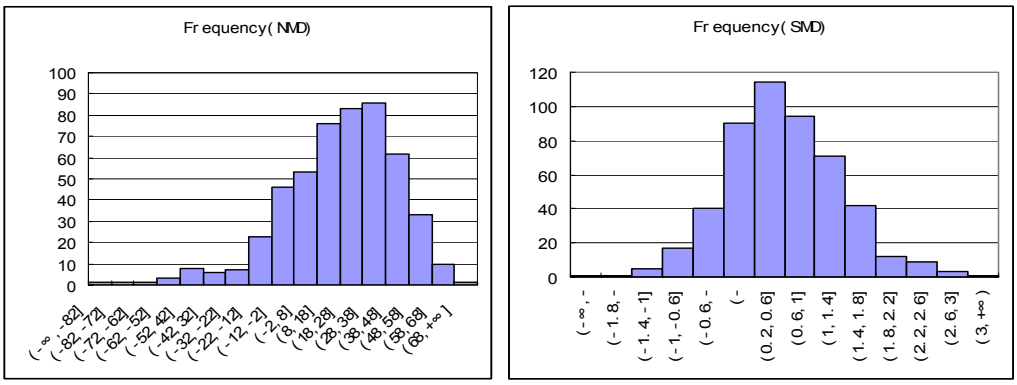


Chart 5.1: Frequency of the effect sizes from 500 single studies in one meta-analysis (a) effect sizes under NMD; (b) effect sizes under SMD

The frequency does not shape too much as the normal distribution. The main reason might be the

small sample size of the single study. The averages of the samples which were used in the calculation of the effect sizes are not unbiased estimates of the means. The powers of the two approaches to calculate the effect size are quite low in this case. Note that the chart of the frequency under SMD shapes more likely than the one under the NMD. That's because the standard errors of effect sizes under SMD are smaller than those under the SMD and a small sample bias adjustment was used in the process of the calculation.

Under the normalised mean difference method, the overall estimate (fixed method) of the efficacy after combining the 500 effect sizes is 18.58282 and its standard error is 0.852901. Hence, the 95% confidence interval is [17.729917, 19.43571989] and only 3% (15 out of 500) of the effect sizes are in this range. For the standardised mean difference method, the overall estimate (fixed method) of the 500 effect sizes is 0.585811 and its standard error is 0.030262 which lead to a range of [0.555549643, 0.616072985] as the 95% confidence interval. There are only 3.6 %(18 out of 500) of the effect sizes that are within this range. When using DerSimonian and Laird random effects models, the overall effect size under NMD method is 18.4552 and its standard error is 1.149799, thus, there are 3.8%(19 out of 500) of the effect sizes are within the range of 95% confidence interval from 17.3054 to 19.605. The overall effect size under SMD method is 0.585368 and its standard error is 0.033855, and there are 4.4 %(22 out of 500) of the effect sizes are within the range of 95% confidence interval from 0.551513 to 0.619223.

The true value of the efficacy is supposed to be 20% improvement, which means the effect size under NMD should be 20 and 0.5 for that under the SMD method. Comparing these to the results from simulation, we can find out that the overall estimate of the efficacy derived from the SMD is overestimated while the one derived from the NMD is underestimated. In addition, the random modules did perform slight better than the fixed modules.

5.2. Power to detect heterogeneity between groups

During the process of the stroke drug test on animals, researchers found that some certain characteristics carried by the research objects might have impacts on the drug. In the paper that report the efficacy of FK506 (a kind of stroke drug being investigated), it wrote[2], "*Study showed*

that the effect size was significantly higher in temporary ischaemia models and in studies using ketamine anaesthesia. Experiments on monkeys gave a higher estimate of effect size than those using rodents, and effect size was higher with healthy animals than where hypertensive or hypolycaemic animals were used." Owing to the issue described as above, it is not proper to give an overall estimate of the drug efficacy only as the assessment of the drug whereas more detailed one is required. The detailed assessment should include the limitation of the effect size of the drug as well. Therefore, it is very important to have a powerful method to detect the heterogeneities between groups.

Besides, some results combined in the meta-analysis are gathered from the published papers, which probably have publishing bias. These published papers usually have the trend to overestimate the efficacy of the drug under investigated whether the researchers are doing it on purpose or not after a few years spent on the drug tests. By using heterogeneity tests, we will have an idea whether the results with publishing bias have a overestimate impact on the overall efficacy when they are combined with the ones do not have the bias. Therefore, a decision can be made if the results in these papers are worth using.

In order to confirm the reliability of the powers of the heterogeneity test derived from the program of the simulation, a check of its correction was done. The check was realized by a comparison of the results calculated by excel versus by the program but with same uniform random numbers. The results obtained in the first stage of the meta-analysis are shown in table5.2. The results in the left two columns are the ones calculated by using the formulae in excel, and the right two columns are the ones output by the java program. Note that there are slight differences in the last few digits of the outcomes. It is because the algorithm of calculating the inversed of the standard normal cumulative distribution used in excel and in the java are different. It cannot tell which one is more accurate as the true value is unknown. Since the errors are less than 10^{-4}, they are all acceptable.

	From excel		From java program	
	control	treatment	control	treatment
Animal 1	65.57883	55.54943	65.57882	55.549421
Animal 2	133.0167	64.47216	133.0168	64.47217
Animal 3	131.8037	114.9706	131.8037	114.97064
Animal 4	109.2888	82.1172	109.2888	82.117211

Animal 5	72.49894	113.6745	72.49893	113.67454
Average	102.4374	86.15679	102.4374	86.156797
S.D.	28.63313	24.53901	28.63314	24.539017
Effect size(NMD)	15.89322133		15.89321699	
SE(NMD)	16.46301033		16.46301386	
Effect size(SMD)	0.551479044		0.551478775	
SE(SMD)	0.651991689		0.651991671	

Table 5.1 data output in the first stage of meta-analysis generating by excel and java program

The results of the comparison in the second stage of meta-analysis are shown in table 5.3. The same summary statistics obtained in the first stage are used in both excel and the java program but not the same as shown above (because the generation of random numbers is not repeatable and the comparison of the two stage were done separately). Only two single studies are combined for the sake of simplicity and clearness. Apparently, the outcomes are exactly the same from both tools.

	Effect size (NMD)	SE(NMD)	Weight (NMD)	Effect size (SMD)	SE(SMD)	Weight (SMD)
Study 1	28.6400286	22.22653	0.002024	0.73608516	0.6668619	2.248683
Study 2	44.3903603	22.97458	0.001895	1.1037418	0.7074711	1.997941
	Overall NMD	SE(NMD)		Overall SMD	SE(SMD)	
from excel	36.25461	15.97444		0.90905932	0.485264	
from java	36.25461	15.97444		0.90905932	0.485264	

Table 5.2 data output in the second stage of meta-analysis generating by excel and java program

The comparison of the heterogeneity is also performed using the two different tools. A hundred theta and their weights, which are divided evenly into two subgroups were used in a comparison. Two comparisons were done under the two approaches of estimating efficacy: NMD and SMD. The heterogeneities between the subgroups are the same calculated by the tools. Hence the power of detecting heterogeneity between groups is convincing. Due to the large scale of the data, the comparisons are not shown in the dissertation.

5.3. Sensitivity analyses of the heterogeneity investigation

Several sensitivity analyses were carried out in order to get a full picture of the power of the stratified meta-analysis. The method of ‘control variable’ was used to see what will influence the power of the meta-analysis. The default values of the sensitivity analyses are shown in the table 5.3. The default difference in the effect sizes between the two subgroups is 10%; the base effect size in the subgroup without characteristic Y is 20%(the mean is 100 in the control group and 80 in the

treatment group); the significant level is 5%; the subgroup size is 20 for either subgroup; and the evenness between the two subgroup is 20 versus 20 (even). Each time, one of the variables changed in a range while the others remain constant as the default value.

effect size	**Base effect size**	**Significant level**	**Single group size**	**Subgroups' sizes**	**Evenness**
10%	100/80	5%	5	20+20	20/20

Table 5.3 Default values of the parameters in the simulation

Since the powers come out from each run of the program varied a little, the simulation program with same parameters was run for ten times (each time the process of the whole simulation was repeated 1000 times) to get a range of the power. The tables and charts in the following content were built on the average of the ten outcomes. Besides, the false positive rate was also tested.

5.3.1. Sensitivity to effect size

First of all, we want to know how slight difference can be detected by the heterogeneity test between the subgroup within which the animals shared the characteristic Y in common and the subgroup that animals have not the characteristic. The results are illustrated in the table 5.4 and chart 5.2. We can see that with the increase of the difference between the subgroups changing from 5% to 50%, the power to detect this difference is increasing as well under either normalised difference mean method or standardised mean difference method. Note that the power of NMD method to detect the difference of effect size is always stronger than that of SMD method and the NMD method can handle slighter difference in effect sizes than the SMD method (NMD is powerful enough to detect 25% difference while SMD is qualified for the 35% difference.). If the difference is tiny such as 5%, the power of SMD approach is only 6.7% while that of NMD approach is 20.4%. The difference in their powers trends to be small, though it increases when the difference in effect sizes change from 5% to 10% and then to 15%. When the difference is large (above 35%), the powers of the two approaches are similar.

EffectSize / Power	*5%*	*10%*	*15%*	*20%*	*25%*	*30%*	*35%*	*40%*	*45%*	*50%*
NMD (%)	20.4	38.9	63.1	82.3	94.5	98.7	99.7	100.0	100.0	100.0

SMD (%)	6.7	17.4	36.7	58.1	78.7	90.7	96.9	99.2	99.9	100.0
Difference	13,7	21.6	26.4	24.3	15.7	8.0	2.8	0.8	0.2	0

Table 5.4: Sensitivity of the powers of the NMD & SMD method to the difference in effect size

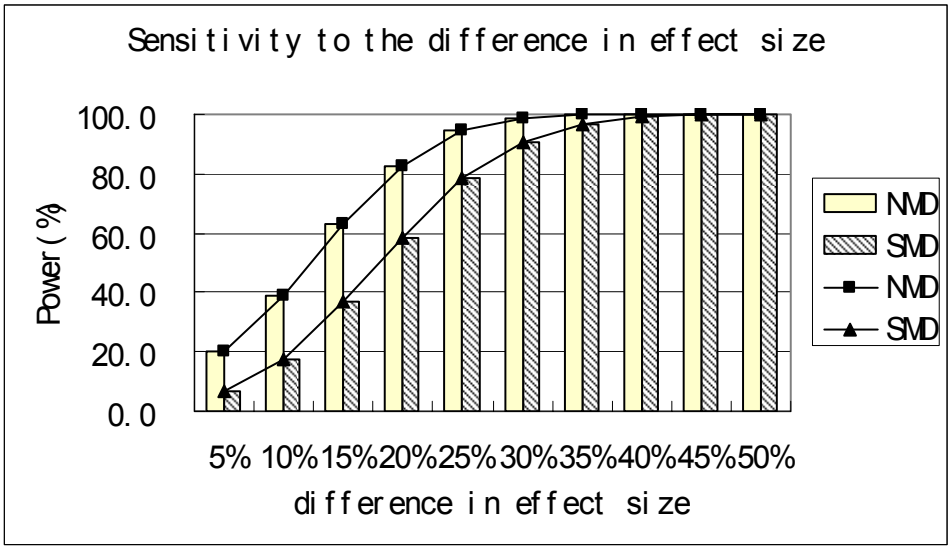


Chart 5.2: Sensitivity to the difference in effect size. The lines between the discrete results are the trend line generate by the polynomial algorithm as the estimate to the ones were not tested.

The results of the sensitive analysis are not object to the common sense. If the impact on the efficacy of the drug brought by the characteristic Y is very large, it will compensate the bias of the small sample and the test will hardly miss the heterogeneity between the subgroups. However, when the difference in effect sizes is small, the data obtained from the small sample are not plenty enough to find out the difference in most cases.

As the power of the tests increase tremendously when the difference of effect sizes between subgroups are getting bigger, it is very sensitive to the difference in effect sizes

5.3.2. Sensitivity to the base effect size

Though the difference in effect sizes between subgroups remain constant, the power to detect heterogeneity might still vary due to the change of the base effect size. The base effect size mentioned here means the overall effect size obtained from the subgroup within which the animals do not have the characteristic Y. For instance, when the base effect size change from 20 %(100/80)

to 10 %(100/90) and the effect size of the other subgroup with characteristic Y change from 30 %(100/70) to 20 %(100/80), the power might change though the difference in effect size maintains 10%. To investigate how the power will change and how sensitive it is, the program of simulation was run with the default parameters shown in table 5.3 but the value of the base effect size was change from 5%(100/95) to 50%(100/50). The results are displayed in the table 5.5 and chart 5.3. There is a slight trend that the power will decrease when the base effect size increase under either NMD or SMD approach.

Base / Efficac / Power	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
NMD (%)	39.65	39.17	38.99	38.16	37.98	37.66	37.38	37.44	36.69	36.65
SMD (%)	18.6	17.43	17.37	16.78	16.73	16.65	16.48	15.72	15.34	14.03

Table 5.5: Sensitivity of the powers of the NMD & SMD method to the base effect size

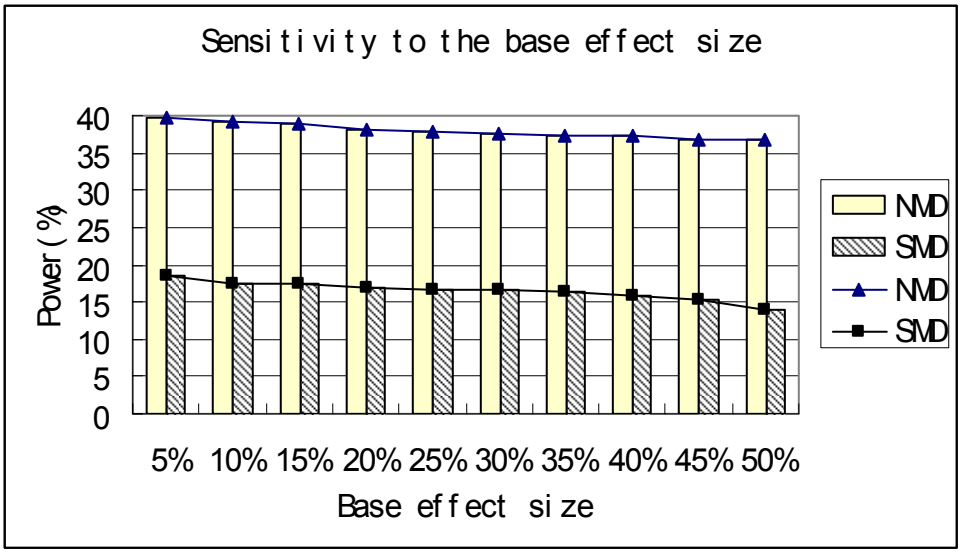


Chart 5.3: Sensitivity to the base effect size. The lines between the discrete results are the trend line generate by the polynomial algorithm as the estimate to the ones were not tested.

It seems to be weird at first glance, however, after having a look at the details and the formulae used in the test, the results begin to make some sense. Observing the details output by the program, the heterogeneity within the subgroups are almost the same while the total heterogeneities are smaller when the base effect size increases. Thus, the observed statistic: the heterogeneity between subgroups is smaller ($Q_B = Q_T - \sum_k Q_k$). This is because when the base effect size becomes bigger

the total difference of all the effect sizes (the summary statistics got from the first stage) is relatively smaller, whereas the differences within each subgroup are unchanged as the effect sizes in each subgroup follows the normal distribution with the same standard deviation (the same shape if drawn as figure). For example, 5 is a big difference between 1 and 6, but not significant different from 195 and 200. Therefore the weights in the formulae $Q_{total} = \sum w_i(\theta_i - \theta_{IV})^2$ become smaller when the base effect size increases.

Since, the power of the heterogeneity test did not change too much when the base effect size increases (only 3% for NMD and 4% for SMD when the base effect size changes from 5% to 50%), we can say it is not very sensitive to the base effect size.

5.3.3. Sensitivity to the significant level

Under different significant level, the behavior of the simulation varies from each other. In the table 5.6, it shows the powers of ten times run of the simulation under 3 significant level 95%, 99% and 99.9%. The power in each column is derived from the same random numbers (the same run of the program). Apparently, the higher the significant level is, the less powerful the test is. In each heterogeneity test, there is only one observed statistic but three critical values in different significant levels which are 3.84146, 6.63489 and 10.82736 for 95%, 99% and 99.9% respectively. When the significant level is 99.9%, the SMD method almost has no power to detect the 10% difference between the subgroups.

		Power of the test (%)										Average
E<0.05	NMD	39.0	37.1	41.0	42.4	39.8	39.6	36.5	40.8	37.4	42.0	39.56
	SMD	18.5	14.4	18.3	17.3	19.2	17.8	15.7	18.9	16.5	20.2	17.68
E<0.01	NMD	23.4	21.2	24.1	24.8	21.7	22.5	20.1	23.5	22.8	24.1	22.82
	SMD	5.8	5.2	5.5	5.3	5.7	5.7	5.4	6.0	4.9	6.1	5.56
E<0.001	NMD	11.6	9.5	10.5	10.4	9.4	10.0	9.4	10.3	8.7	11.5	10.13
	SMD	0.9	0.9	0.7	0.7	0.8	0.5	0.5	0.8	1.3	0.9	0.80

Table 5.6: The powers of the heterogeneity test under different significant level (10 times)

There seems to be no direct relationship among the three levels, the increase of power in one level does not mean the same increase in another level. Since the power is too small in the level of 99% and 99.9%, it is better to use the 95% significant level if the precision of the test is not highly

required.

5.3.4. Sensitivity to the individual numbers of a single study

The most important purpose of the project is to find out which method is more proper to use when the sample size of the single study is small. The simulation was run from a small sample size of five to a relatively large sample size of 100. Besides the sensitivities of the powers of the two approaches were tested, the sensitivities of false positive rate were tested as well. A ROC (chart 5.4) curve was drawn to show the relationship of the power and the false positive rate of the heterogeneity tests. With the increase of the sample size, the powers of both NMD and SMD method grow rapidly. The power of the NMD increases from 38.8% (at 5 animals per group) to 100% (at 100 animals per group), while the power of the SMD increases from 17.7% (at 5 animals per group) to 100%(at 100 animals per group). The power of NMD method is always greater than that of the SMD.

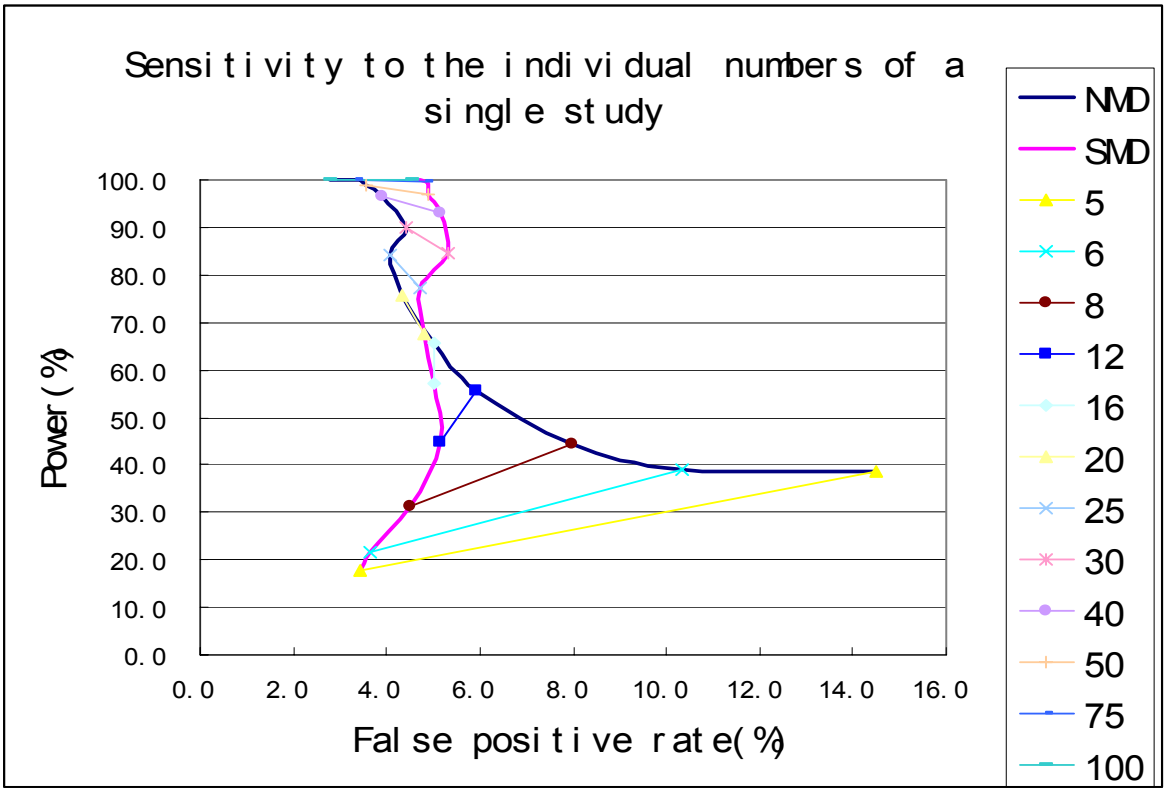


Chart 5.4: The Roc curve of the sensitivity to the individual numbers of a single study.

Power

	5	**6**	**8**	**12**	**16**	**20**	**25**
NMD(%)	38.8	38.8	44.4	55.5	65.8	75.7	84.2

SMD(%)	17.7	21.8	31.3	44.8	57.2	67.8	77.2
False Positive Rate							
NMD(%)	14.5	10.4	8.0	5.9	5.0	4.4	4.1
SMD(%)	3.4	3.7	4.5	5.2	5.0	4.8	4.7

Table 5.7 (A): the data of the sensitivity to the single group's size

Power					
	30	**40**	**50**	**75**	**100**
NMD(%)	90.1	96.5	98.9	100.0	100.0
SMD(%)	84.4	93.1	97.1	99.8	100.0
False Positive Rate					
NMD(%)	4.4	3.9	3.6	3.3	2.8
SMD(%)	5.3	5.2	4.9	4.9	4.6

Table 5.7(B): the data of the sensitivity to the single group's size (continued)

The false positive rates were tested by setting the difference in the effect size as zero between the two subgroups. Unlike the powers, the false positive rates of the NMD approach decrease while that of SMD approaches increase when the sample size of the single study becomes bigger. The slopes of the lines that link the powers and false positive rates are getting bigger and rotate anti-clockwise. When the sample size is larger than 16 animals per group, the false positive rates under SMD are greater than that under NMD. To explain this, we should know that the total error of the test consists of the type□error and type□error. The value of the power equals to one minus type□error and the false positive is type□error. When the number of the single study increase, the total error of the test are smaller. For the NMD approach, the ratio of type□error decrease and the ratio of type□error increase relatively but the value of both the type I error and type□error become smaller. For the SMD approach, the ratio and of type□error increase while the value and ration of the type I error decrease. For example, suppose the total error is 8%, the type I error is 5% and the type□error is 3% when the sample size is 5. When the sample size changes to 10, the total error will reduce to 7%, the type□error will increase to 4% and the type I error will decrease to 3%, which means the power will be 97% and the false positive rate will be 4% (These numbers are not the results come from the simulation). Note that the false positive rate is not mono-decrease or mono-increase because its value is quite small, a little noise will influence it easily.

From the sensitive analysis, we can see that the SMD approach is less powerful than the NMD approach when the sample size of the single study is small. However, the NMD approach might

overestimate the heterogeneity between the groups.

5.3.5. Sensitivity to the subgroups' sizes

Besides the change of the numbers of animals in a single group, the other way to increase the total numbers of animals used in a research is to increase the number of single studies in one combination. Again, the program of the simulation was run several times to see how the power of the two approaches would change.

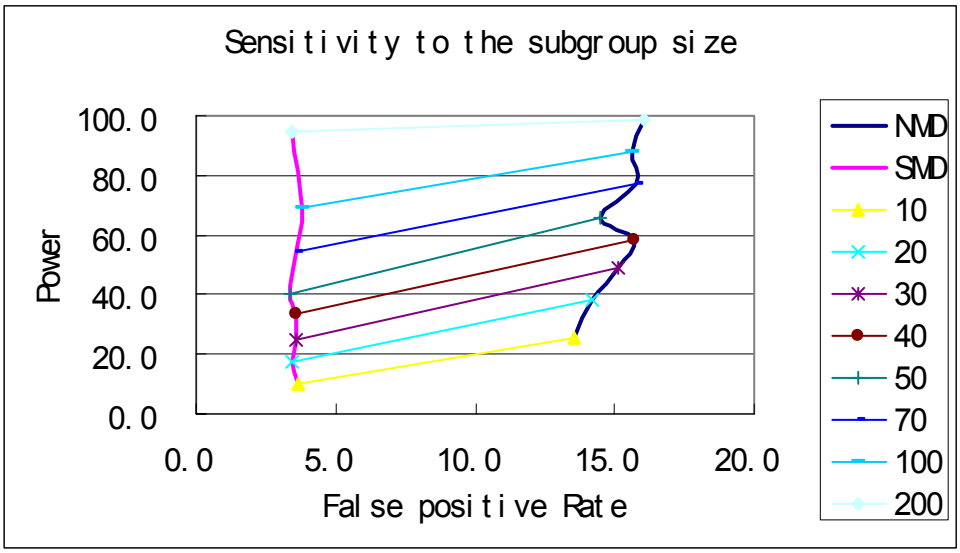


Chart 5.5: The Roc curve of the sensitivity to the subgroups' sample size

Sensitivity to the subgroups' sizes				
Subgroups' size	Power (%)		False Positive Rate (%)	
	NMD	SMD	NMD	SMD
10+10	**25.6**	**10.3**	**13.6**	**3.7**
20+20	**38.4**	**17.3**	**14.2**	**3.4**
30+30	**49.0**	**25.1**	**15.1**	**3.6**
40+40	**58.7**	**33.4**	**15.7**	**3.6**
50+50	**65.7**	**40.3**	**14.5**	**3.4**
70+70	**77.3**	**54.3**	**15.8**	**3.6**
100+100	**88.1**	**69.1**	**15.6**	**3.8**
200+200	**98.7**	**94.8**	**16.0**	**3.5**

Table 5.7: the data of the sensitivity to the subgroup groups' size

In the chart 5.5, it shows the ROC curve of the powers and false positive rate of NMD method and SMD method. The detailed the data is showing in the table 5.7. Both of the powers increase when the numbers of single studies combined in one subgroup become larger. However, the powers of both approaches do not seem to be promising even the subgroup's size is 100 (200 in total). The

performance of the NMD approach is better than that of SMD under all sample sizes. However, the false positive rate for NMD is incredible higher than that of SMD. It seems that there is not any particular rule of the change of the false positive rate under either NMD approach or SMD approach. It can be believe that the false positive rate remains a constant, and the variety of the values shown in the table 5.7 is the play of the chance. Or the underlying relationship between the false positive rate and the subgroups' sample size is slight enough to be ignored. Note that even though there are 200 single studies in each subgroup, the power of either NMD or SMD can achieve 100 %(1000 out of 1000 times the heterogeneity test can detect the difference in effect size between two groups). The possible explanation is that the bias caused by small number of animals in each single study still has some influence on the test though the total number of animals is already very large.

5.3.6. Sensitivity to the evenness of the subgroups' sizes

In practice, the numbers of the single studies combined in subgroups are not exactly the same. The unevenness of the subgroup might influence the power of the heterogeneity test. The interest is under which circumstances (even or uneven), the test is more powerful. Set the total number of the single studies as a constant (40 here), and then run the simulation with different allocation of single studies between the two subgroups. For instance, derive the power of the heterogeneity test with 2 single studies in the subgroup without Y and the rest 38 in the subgroup. Then increase the number of single studies in the subgroup without Y to 4 and decrease that in the subgroup with Y to 36. Go on with this until the number of single studies in the subgroup without Y is 38 and that in the other subgroup is 2. The detailed results are shown in Table 5.8 and illustrated in chart 5.6.

Observing the curve of the power in chart 5.6, it can be considered as concave. The power achieves the peak when the single studies are divided into two subgroups evenly and performs poorest at both ends, when the numbers of the single studies in the subgroups are most not balanced. The NMD method seems to be more sensitive to the evenness between subgroups than the SMD method as the curve of NMD is steeper than that of the SMD.

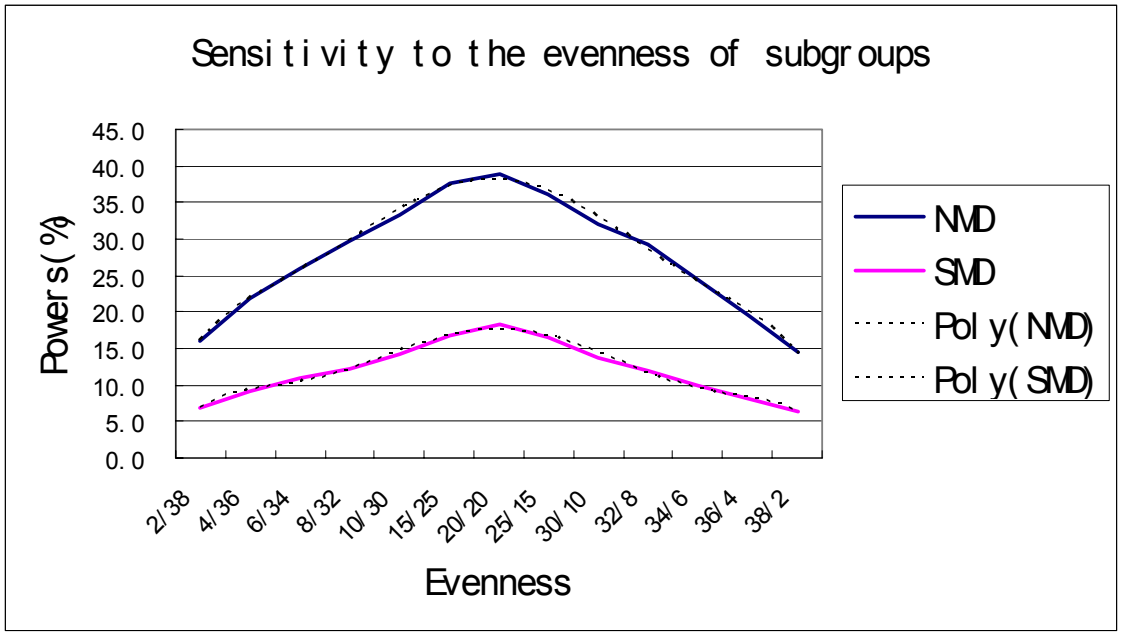


Chart 5.6: The powers of NMD & SMD with different evenness of subgroups' sizes

	2/38	**4/36**	**6/34**	**8/32**	**10/30**	**15/25**	**20/20**
NMD(%)	16.1	21.9	25.9	29.8	33.4	37.6	38.9
SMD(%)	6.8	9.1	10.9	12.1	14.1	16.8	18.3

Table 5.8(A): Data of the sensitivity analysis to the evenness of subgroups'

	25/15	**30/10**	**32/8**	**34/6**	**36/4**	**38/2**
NMD(%)	36.2	32.0	29.4	24.5	19.6	14.4
SMD(%)	16.6	13.8	12.0	9.9	8.0	6.4

Table 5.8(B): Data of the sensitivity analysis to the evenness of subgroups'

We know that the heterogeneity depends on the number of the studies. When the subgroup sizes are not balanced, there must be a subgroup has fewer studies to combine than the other. The heterogeneity within the larger group is much greater than the one if the numbers are balanced and the heterogeneity within the smaller group is not smaller enough to reduce the increase amount, hence, the sum of the heterogeneity within groups become larger compared to the one when the subgroups are even. Since the total number of the study is unchanged, the total heterogeneity can be regarded as a constant. Therefore, as the difference of the total heterogeneity and the within groups heterogeneities, the heterogeneity between groups become smaller and it is more difficult to be detected. For example, in one simulation, when there are 2 studies in the subgroup without Y and 38 studies in the subgroup with Y, the heterogeneities within group are 0.33794 and 45.63919 respectively and the total heterogeneity is 46.09225. When the numbers of the subgroup are even,

the heterogeneities within group are 24.45490 and 17.99991 for the subgroup without Y and with Y respectively and the total heterogeneity is 42.77561.

According to the results when a research is prepared, it is better to divide the studies evenly so that more promising outcomes can be obtain with same or less amounts of animals

6. Conclusion & Discussion

6.1. Conclusion

Using meta-analysis to combine the outcomes from the animal drug investigation is a work should be undertaken with caution. Due to the bias brought by the small sample size, the meta-analysis seems to lose its power, unlike when it is applied in the clinical trials. The sufficient way to improve the credibility of the estimate of the drug efficacy is to increase the sample size.

It seems that the normalised mean difference approach has a better performance than the standardised mean difference approach. First, the NMD is more powerful than the SMD under most circumstances, in particular, when the sample size is small. In other words, fewer animals are involved in the investigation to gain the same promising outcome by using the NMD method than using the SMD method. Secondly, the overall estimate of drug efficacy derived from NMD approach is straighter and easier to interpret than the one obtained from SMD approach. The outcome derived from NMD is the normalised mean in difference, usually expressed as the percentage of the improvement of the mean, for example, 30%, and it can be regarded the efficacy of the drug directly. The result derived from SMD is the outcome after standardization, usually expressed as the units of the standard deviation rather than in units of any of the measurement scales used in the research, for instance, 0.8s.d will appear in the report as the effect size of the new drug. It cannot tell the efficacy directly as no one knows how much is one standard deviation stand for. The way to get the understandable efficacy for SMD method is to do the whole standardised process backwards. Finally, NMD method can handle the data measuring the same underlying effect but using different scales as the same as the SMD method does. The process of normalizing can turn all the data to a uniform scale. However, the NMD approach might have a higher probability to overestimate the heterogeneity between subgroups as its false positive rate is higher than that of the SMD approach.

When the difference in effect size between the two subgroups is larger, it is more powerful of the

heterogeneity investigation. It is a contrary situation when the base effect size becomes bigger, but the decrease trend of power seems to be very slight and when the sample size of the study is big enough the trend can be ignored.

Not only the individual number of animals used in a single study but also the numbers of single study combined in a subgroup can influence the power to detect the heterogeneity between the subgroups. The more animals involved in the research, the more powerful the heterogeneity test is. Besides, the evenness between subgroups has the impact on the power of the heterogeneity test. For the same amounts of total single studies in the stratified meta-analysis, the most powerful allocation way is to divide the single studies to the subgroups evenly.

As for the significant levels, the heterogeneity test is most powerful at 95% compared to 99% and 99.9% since less precision is required. The suggestion is to use 95% significant level as possible as it can otherwise the meta-analysis will hardly have any power when the sample size is small.

6.2. Discussion

The results generated by the simulation show a reasonable picture of how the stratified meta-analysis performs using NMD and SMD method in the animal drug investigation. However, there are still some potential difficulties with the simulation. First, due to the random nature of the simulation, the final outcomes (the powers) generated varied from time to time even though the run times of the simulation process is 10000. The powers are just a rough estimate of the true value. We cannot regard the powers come out in one run of the simulation program as the final answer. By running the program several times, we can only learn about a range within which the true value is. Secondly, the heterogeneity is a non-randomized comparison. The noise brought by the small sample can either reduce the power of test but also increase it. As shown in previous chapter, the NMD approach is proved to be more powerful than the SMD approach by the simulation, but its high false positive rate is still an issue to concern. The power of NMD approach derived in the simulation is probably overestimated. Thirdly, the program of the simulation can only compared two subgroups, which is not enough in practice. In the practical research, there are probably more than one characteristic will impact the efficacy of the drug. It is insufficient and improper to

compare each subgroup one by one. Further work can be done to the program so that the user can choose the number of subgroups they want and do the heterogeneity investigation between these subgroups. Fourthly, the program of the simulation is a black box to the user and cannot supply any detailed underlying information. It is not convenient when the user want to know the overall estimate of the effect size and its confidence interval. The observed heterogeneity statistics are also unknown to the user. Additional function can be added to the program to show the details and whether or not to display the outcomes. Finally, the simulation can not deal with the sample size below 2 (included) of a single study due to the formulae used (see chapter 2.2).

References

1 Deeks JJ. Systematic reviews of published evidence: miracles or minefields? *Ann Oncol* 1998; **9:**703-9.

2 Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and meta-analysis of the efficacy of FK506 in experimental stroke *Journal of Cerebral Blood Flow & Metabolism* 2005; 1-9.

3 Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis *System Reviews in Health Care*, 2nd edn. London: BMJ Books, 2001: 286-312

4 Hedges LV, Olkin I. Parametric Estimation of Effect Size from a Series of Experiments *Statistical Methods for Meta-Analysis* Orlando, Florida Academic Press, 1985: chapter 6

5 Sinclair JC, Bracken MB. *Effective care of the newborn infant*. Oxford: Oxford University Press, 1992: chapter 2

6 Rosenthal R. Parametric measures of effect size. In: Cooper H, Hedges LV, eds. *The Handbook of research synthesis.* New York: Russell Sage Foundation, 1994.

7 DerSimonian R, Laird N. Meta-Analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177-88.

8 Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, 2001;**20**: 825-840

9 B□hning D, Malzahn U, Dietz E, Schlattmann P. Some general points in estimatin heterogeneity variance with the DerSimonian-Lair estimator *Biostatistics* 2002; **3**: 445-457

Appendix A

Sensitivity to the difference in effect size											
	Power										mean
5%	18.2	19.5	19.7	23.8	19.7	20.5	20.0	20.7	21.1	20.4	20.4
5%	5.2	7.0	5.8	7.8	7.1	6.5	6.0	7.2	7.1	7.1	6.7
10%	37.2	38.9	39.3	40.8	39.0	40.0	40.4	36.4	39.7	37.7	38.9
10%	18.0	16.9	17.4	16.6	18.8	16.7	17.0	19.3	18.0	15.0	17.4
15%	65.2	61.2	64.4	65.6	63.3	61.6	63.2	60.9	61.1	64.6	63.1
15%	38.7	35.9	36.2	38.9	37.0	36.2	37.3	35.8	36.2	34.6	36.7
20%	83.8	81.0	83.7	80.2	81.3	82.1	83.4	81.9	84.1	81.9	82.3
20%	59.4	57.7	58.2	56.1	57.8	55.6	59.3	58.6	60.9	57.0	58.1
25%	94.5	95.1	93.7	94.2	93.5	94.8	94.8	95.1	95.0	93.9	94.5
25%	77.8	79.1	77.3	80.7	78.3	78.7	80.7	79.6	78.9	76.3	78.7
30%	98.3	98.1	98.4	99.5	99.3	98.4	98.8	98.9	98.5	98.4	98.7
30%	89.2	91.1	90.1	90.5	91.0	91.9	90.3	91.6	90.8	90.2	90.7
35%	99.4	99.8	99.7	99.9	99.6	99.7	99.7	99.7	99.8	99.7	99.7
35%	96.9	97.0	96.5	97.5	96.5	97.3	96.1	97.2	96.8	97.1	96.9
40%	100	100	99.8	99.9	100	100	100	99.9	100	100	100
40%	98.9	99.4	99.0	99.3	99.3	99.7	99.2	98.7	98.9	99.5	99.2
45%	100	100	100	100	100	100	100	100	100	100	100
45%	99.8	99.8	99.7	99.9	99.8	99.9	100	99.9	99.8	99.9	99.9
50%	100	100	100	100	100	100	100	100	100	100	100
50%	99.9	100	99.9	100	100	99.9	100	100	100	100	100

c_mean = 100
t_mean_withoutY=80
Pop_sd = 40
n_control=5
n_treatment=5
Total Experiments_No.without Y = 20
Total Experiments_No.with Y = 20
normlised(power)
standardised(power)

Appendix B

Sensitivity to the base effect size											
T_mean withoutY	Power										mean
95%	40.1	41.1	41.2	37.0	39.6	39.2	38.7	38.7	40.6	40.3	39.7
95%	19.8	18.9	18.2	18.4	18.5	18.4	17.9	19.0	18.8	18.1	18.6
90%	37.9	39.9	39.1	38.3	37.4	39.6	39.7	41.0	39.6	39.2	39.2
90%	15.9	17.8	17.9	17.9	15.8	19.5	17.3	17.9	17.9	16.4	17.4
85%	39.4	37.2	39.0	38.6	36.6	39.3	39.1	40.1	39.7	40.9	39.0
85%	19.6	15.9	17.4	17.4	16.9	17.9	16.5	18.5	16.8	16.8	17.4
80%	37.0	38.8	37.9	39.4	38.5	37.1	38.9	36.9	38.4	38.7	38.2
80%	15.7	18.3	16.7	16.8	15.7	16.8	16.9	16.8	15.6	18.5	16.8
75%	36.7	38.5	38.7	38.7	38.1	39.6	37.8	37.3	36.4	38.0	38.0
75%	16.8	18.9	16.2	16.8	16.0	16.3	17.1	16.6	15.8	16.8	16.7
70%	37.9	35.4	38.9	38.5	38.5	37.7	37.0	38.1	37.8	36.8	37.7
70%	16.2	16.6	17.2	17.6	17.2	16.0	16.9	17.3	16.5	15.0	16.7
65%	34.5	37.7	38.6	37.7	35.8	38.4	38.0	38.7	37.6	36.8	37.4
65%	16.4	15.9	17.3	15.7	16.1	16.7	15.8	17.4	17.8	15.7	16.5
60%	36.4	35.9	37.8	37.2	38.8	37.9	39.4	35.4	40.1	35.5	37.4
60%	13.2	15.4	14.5	15.5	16.9	14.3	17.4	17.1	17.9	15.0	15.7
55%	35.9	37.1	37.5	36.1	35.7	35.7	36.2	38.1	37.3	37.3	36.7
55%	15.4	15.3	13.5	16.4	14.2	15.3	15.5	15.5	16.0	16.3	15.3
50%	34.3	35.7	36.6	37.1	35.8	37.2	39.3	38.6	35.4	36.5	36.7
50%	14.2	12.6	12.1	15.2	13.8	13.5	15.9	13.4	16.5	13.1	14.0

c_mean = 100
Pop_sd = 40
n_control=5
n_treatment=5
Total Experiments_No.withoutY = 20
Total Experiments_No.withY = 20
normlised(power)
standardised(power)

Appendix C

Sensitivity to the No. of individual studies at 5%											
	Power										mean
5	38.4	38.0	38.6	40.1	41.0	40.3	36.9	38.5	39.3	36.8	38.8
5	18.7	18.3	17.3	18.7	19.7	19.0	14.5	16.3	18.5	15.7	17.7
6	38.6	41.0	40.6	37.9	39.4	39.3	38.0	39.2	36.6	37.5	38.8
6	21.8	23.4	22.9	21.4	23.0	21.3	20.7	21.4	22.2	19.4	21.8
7	40.1	42.9	41.7	42.8	41.6	44.3	42.6	43.7	44.3	43.5	42.8
7	25.5	28.5	26.9	27.8	25.6	28.2	26.8	27.5	27.2	27.9	27.2
8	48.5	44.9	43.0	45.2	43.0	44.2	44.4	45.1	44.6	41.3	44.4
8	33.4	30.2	30.3	31.1	30.9	31.4	30.8	32.2	31.6	31.2	31.3
9	47.7	48.0	45.5	47.8	45.4	46.1	46.3	43.5	46.0	48.3	46.5
9	35.6	33.5	32.4	35.7	32.2	35.3	38.6	31.7	33.2	34.9	34.3
10	48.1	48.3	51.3	46.6	48.7	47.7	50.1	47.2	50.2	49.2	48.7
10	37.9	36.9	38.5	36.2	37.3	37.5	38.6	37.1	40.2	37.8	37.8
12	55.9	54.0	56.1	54.0	52.5	56.3	56.2	57.1	54.4	58.3	55.5
12	45.5	42.7	45.9	44.2	42.4	45.6	44.7	46.6	44.5	46.3	44.8
14	60.2	59.7	62.5	59.7	63.0	65.7	60.5	61.8	59.7	62.1	61.5
14	51.5	50.7	52.1	50.4	53.3	54.1	52.0	52.5	49.9	51.1	51.8
16	66.6	64.8	67.0	66.1	66.3	66.7	65.5	64.6	65.4	64.8	65.8
16	59.7	55.8	56.1	56.8	56.5	58.6	58.4	57.4	55.8	56.6	57.2
18	70.6	71.8	71.6	70.7	71.6	71.5	72.3	74.1	70.1	70.9	71.5
18	61.8	62.3	62.5	61.8	64.1	61.8	61.3	66.9	61.6	60.8	62.5
20	77.0	73.9	77.5	75.6	74.0	74.4	73.2	77.1	77.9	76.0	75.7
20	69.0	67.6	70.1	67.4	65.2	67.2	66.1	68.9	68.5	67.6	67.8
25	83.9	85.3	86.2	83.3	84.7	83.3	84.2	84.8	83.4	83.1	84.2
25	76.9	78.3	80.1	75.7	76.9	76.0	77.6	77.7	76.9	75.5	77.2
30	89.4	90.9	89.7	90.8	89.9	90.0	88.2	91.2	90.3	90.2	90.1
30	83.9	84.8	83.2	84.3	83.1	84.2	84.4	86.5	84.6	85.1	84.4
40	97.1	96.3	96.3	94.9	97.4	96.5	97.0	96.1	96.7	96.7	96.5
40	93.6	92.9	93.0	92.2	93.8	93.3	93.5	92.0	93.4	92.8	93.1
50	98.9	99.1	99.7	98.6	99.0	99.0	98.4	98.3	98.3	99.2	98.9
50	96.6	96.5	98.0	96.8	97.2	97.9	97.2	96.4	96.5	97.7	97.1
75	100	99.9	100	100	99.8	100	99.9	100	100	99.9	100
75	99.9	99.8	99.7	99.7	99.6	99.8	99.8	99.9	99.9	99.8	99.8
100	100	100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100	100	100

	False Positive at 5%										mean
5	14.3	15.3	13.6	14.3	14.7	16.2	12.4	15.6	15.0	13.6	14.5
5	3.8	3.3	3.1	3.8	3.7	3.7	2.3	4.0	3.3	3.4	3.4
6	10.7	11.1	9.4	12.2	11.0	10.8	7.6	11.9	8.9	9.9	10.4
6	3.5	3.5	4.7	4.2	3.9	3.1	3.1	4.0	3.0	3.6	3.7
7	8.6	10.1	8.9	10.8	8.7	10.1	10.2	10.0	10.8	9.5	9.8
7	3.9	5.2	4.2	4.4	4.3	5.7	5.4	3.8	5.5	5.1	4.8
8	7.2	6.9	8.1	8.6	8.4	8.2	6.6	8.0	9.6	8.1	8.0
8	4.2	3.1	4.1	5.1	4.8	4.8	3.8	5.7	4.7	4.8	4.5
9	8.0	7.3	8.1	6.5	7.8	8.3	6.8	7.8	7.1	6.9	7.5
9	5.0	4.2	5.1	4.3	5.1	5.0	3.9	4.4	4.0	3.7	4.5
10	6.6	6.3	7.8	5.2	6.3	7.6	6.2	6.1	6.9	6.5	6.6
10	4.8	3.7	4.4	3.7	4.6	5.3	4.1	4.2	4.3	4.6	4.4
12	6.3	5.2	5.5	6.1	6.1	5.6	6.1	7.0	5.9	5.6	5.9
12	5.9	4.7	4.1	5.1	5.5	4.9	5.7	5.8	5.2	4.7	5.2
14	5.5	4.4	4.6	5.3	5.4	6.4	4.8	5.9	6.0	4.0	5.2
14	4.8	4.7	4.3	4.7	4.8	6.0	4.7	5.2	5.0	4.6	4.9
16	5.8	4.7	4.9	4.7	5.2	6.3	4.5	5.3	5.1	3.5	5.0
16	4.4	4.4	5.5	4.7	5.2	7.2	4.7	5.0	5.0	4.3	5.0
18	3.9	3.7	4.7	2.9	5.1	5.0	5.3	4.0	4.6	5.2	4.4
18	5.0	4.7	5.4	4.1	5.0	4.9	6.0	4.4	4.6	5.2	4.9
20	4.4	4.6	4.0	4.3	3.9	3.8	4.7	3.7	4.9	5.2	4.4
20	4.6	4.8	4.7	5.1	3.8	4.6	4.8	4.0	5.6	6.1	4.8
25	3.7	4.7	4.4	3.8	3.6	3.9	3.3	4.9	4.8	3.5	4.1
25	4.5	4.8	5.2	4.5	4.7	4.5	4.3	5.4	5.0	4.2	4.7
30	4.8	3.8	4.7	5.5	3.0	4.0	3.9	5.6	5.0	3.8	4.4
30	5.8	4.9	5.8	6.3	3.5	5.6	4.3	6.4	5.9	4.7	5.3
40	3.5	4.3	4.2	3.6	3.0	4.5	5.5	3.5	3.6	3.5	3.9
40	5.6	6.1	4.6	5.1	5.0	5.9	6.1	3.8	4.6	4.8	5.2
50	4.6	3.5	3.9	3.5	3.1	2.9	3.7	3.6	3.6	3.1	3.6
50	5.8	4.9	4.6	5.0	4.2	5.3	5.1	5.3	4.6	4.3	4.9
75	2.7	3.5	3.8	3.2	2.8	4.7	3.4	2.1	3.9	3.2	3.3
75	3.9	5.5	5.0	4.3	4.4	6.0	5.4	3.7	5.7	4.6	4.9
100	3.3	2.5	2.6	3.3	2.8	2.5	2.6	2.5	2.2	3.7	2.8
100	5.6	5.0	4.4	5.5	4.2	4.0	4.2	3.8	4.4	4.5	4.6

c_mean = 100	Total Experiments_No.withY = 20
t_mean_without Y = 80	normlised(power)
t_mean_with Y = 70	standardised(power)
Pop_sd = 40	
Total Experiments_No.withoutY = 20	

Appendix D

Sensitivity to the No. of Experiments at 5%											
	Power										mean
10	26.8	23.4	26.3	23.8	26.3	24.2	23.2	27.0	28.1	26.5	25.6
10	10.2	9.5	9.3	9.3	11.1	11.1	9.7	11.6	11.5	9.8	10.3
20	38.8	37.5	39.4	37.6	37.6	37.4	38.1	40.4	37.3	40.0	38.4
20	17.9	16.4	18.2	18.1	16.0	16.2	18.3	16.5	16.9	18.3	17.3
30	47.1	52.7	49.0	49.8	49.1	49.0	49.4	48.3	47.6	47.8	49.0
30	25.1	28.8	24.3	24.3	25.2	25.5	24.7	25.0	24.3	24.1	25.1
40	56.9	56.7	58.8	58.5	60.4	55.9	60.8	62.3	57.1	59.8	58.7
40	34.1	31.8	33.6	34.9	36.4	30.3	33.6	34.0	30.9	34.1	33.4
50	65.4	66.0	63.3	66.8	67.1	66.4	64.9	66.4	66.1	64.8	65.7
50	39.0	39.6	39.1	42.7	39.6	42.0	38.6	41.4	41.1	39.5	40.3
70	80.0	76.3	78.0	79.5	77.0	75.4	77.5	75.1	76.0	78.1	77.3
70	57.1	53.6	55.1	54.8	56.3	51.9	54.0	54.7	52.0	53.1	54.3
100	87.8	89.1	89.6	87.3	87.4	88.3	88.3	87.7	88.2	87.5	88.1
100	70.4	68.1	71.5	67.7	70.9	69.7	67.2	69.6	68.2	67.8	69.1
200	98.9	99.2	98.7	98.9	99.0	98.1	98.7	98.4	98.5	98.8	98.7
200	94.9	95.6	94.7	94.6	95.2	94.6	95.0	94.5	94.4	94.0	94.8
	False positive										mean
10	12.7	13.8	12.4	15.1	12.5	14.5	11.8	14.6	14.9	13.4	13.6
10	3.5	3.6	2.3	4.6	3.3	3.4	3.7	3.7	4.8	3.9	3.7
20	12.0	14.5	15.2	14.4	14.7	13.9	14.2	14.1	15.0	13.7	14.2
20	3.2	4.0	4.0	3.6	3.7	3.0	4.2	2.8	2.6	3.2	3.4
30	15.3	15.8	14.8	13.5	14.0	15.3	13.6	15.6	16.2	16.8	15.1
30	3.1	2.9	4.0	3.1	4.3	3.8	2.9	3.2	4.3	4.2	3.6
40	15.4	14.7	16.6	14.9	15.5	18.8	14.8	13.9	15.4	17.2	15.7
40	3.6	3.2	4.2	4.0	3.8	4.3	3.1	4.1	2.6	2.6	3.6
50	15.6	14.6	13.0	13.8	15.5	16.0	14.6	14.2	14.0	13.6	14.5
50	3.9	4.5	3.2	3.4	2.8	3.4	3.5	3.6	2.3	3.2	3.4
70	16.0	13.6	15.8	15.7	17.4	16.4	16.7	15.9	15.2	15.3	15.8
70	3.1	3.0	3.8	3.5	4.4	4.2	2.4	4.1	4.0	3.3	3.6
100	16.0	15.7	15.0	15.4	15.5	16.1	14.8	15.1	15.5	17.3	15.6
100	3.9	3.1	4.5	3.4	3.1	4.7	4.1	4.0	3.5	3.9	3.8
200	14.9	16.2	16.1	17.7	16.7	15.1	16.1	15.3	16.6	15.7	16.0
200	3.0	4.3	4.2	3.2	4.2	2.5	2.7	3.3	4.0	3.1	3.5

c_mean = 100	n_control=5
t_mean_without Y = 80	n_treatment=5
t_mean_with Y = 70	normlised(power)
Pop_sd = 40	standardised(power)

Appendix E

Sensitivity to the evenness of subgroup											
	Power										mean
2/38	17.7	16.7	16.5	14.3	15.6	17.3	16.8	14.5	16.0	15.2	16.1
2/38	7.6	8.0	6.8	6.5	6.7	7.0	7.3	6.2	6.4	5.7	6.8
4/36	22.8	20.9	24.5	20.5	22.7	20.1	21.9	21.4	21.5	22.3	21.9
4/36	8.0	8.0	9.8	9.5	11.4	8.6	9.4	9.0	8.1	8.8	9.1
6/34	26.0	27.7	26.3	29.4	25.4	24.8	24.4	24.8	26.0	23.8	25.9
6/34	10.6	11.3	11.1	13.2	10.7	10.4	9.5	10.4	10.8	11.3	10.9
8/32	32.2	29.8	28.8	30.1	28.2	33.2	29.5	29.6	26.2	30.5	29.8
8/32	11.6	12.0	11.4	10.8	11.1	14.5	11.6	13.1	11.3	13.6	12.1
10/30	33.4	32.6	32.5	37.8	32.7	32.7	32.9	32.9	33.9	32.9	33.4
10/30	15.0	11.9	15.0	16.1	12.6	14.2	14.5	14.3	13.9	13.7	14.1
15/25	36.9	36.2	39.7	39.5	37.1	36.0	36.5	37.4	37.0	40.1	37.6
15/25	17.7	16.6	16.1	19.4	18.7	14.6	16.7	15.0	16.6	16.6	16.8
20/20	40.3	39.9	38.2	37.2	37.3	40.0	38.2	41.3	37.1	39.6	38.9
20/20	19.7	19.4	16.2	17.2	19.1	18.8	18.0	17.1	18.2	19.1	18.3
25/15	33.1	36.7	34.3	38.4	36.3	36.8	35.8	37.6	36.2	37.0	36.2
25/15	16.1	16.6	15.8	18.3	16.8	17.7	15.4	16.7	16.9	15.4	16.6
30/10	30.5	30.9	32.8	32.1	33.8	34.6	31.7	29.8	29.5	33.8	32.0
30/10	12.7	13.4	14.0	13.8	14.8	14.2	14.0	13.3	12.8	15.0	13.8
32/8	30.7	28.8	29.9	27.5	27.4	28.8	33.2	28.5	29.9	28.9	29.4
32/8	12.3	10.5	13.2	10.9	12.5	11.6	13.2	11.3	11.8	12.2	12.0
34/6	21.6	22.9	23.0	25.4	24.9	25.0	24.8	24.7	25.6	26.8	24.5
34/6	9.0	9.2	8.5	9.9	11.2	10.3	9.9	10.6	9.7	10.6	9.9
36/4	19.4	21.9	19.9	19.7	17.9	21.2	19.6	19.9	19.5	17.3	19.6
36/4	9.5	9.0	8.7	7.1	7.8	7.3	7.7	8.9	7.5	6.8	8.0
38/2	14.4	15.5	14.6	15.5	12.8	16.2	12.6	15.8	14.1	12.7	14.4
38/2	6.4	6.6	6.5	6.4	5.9	6.2	6.6	6.6	6.6	6.4	6.4

c_mean = 100	n_control=5
t_mean_without Y = 80	n_treatment=5
t_mean_with Y = 70	normlised(power)
Pop_sd = 40	standardised(power)
Total number of single studies = 40	