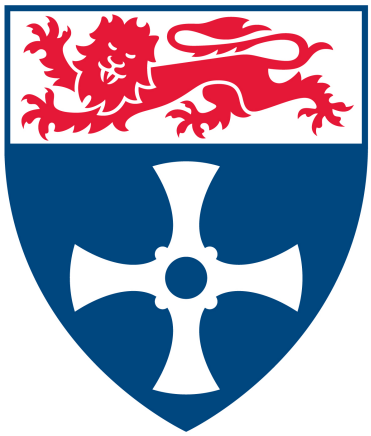

Newcastle University

MAS8391: MMathStat Project

School of Mathematics & Statistics

Graphical techniques to detect and adjust for heterogeneity in meta-analysis

Author:
Karl Delargy

Supervisor:
Dr. Peter Avery

April 30, 2015

Abstract

Meta-analysis techniques, comprising of both continuous and binary data, to detect and adjust for heterogeneity are applied to both simulated and real life datasets. The distinction between the two key models, fixed effects model and random effects model, will be explored. Fundamental concepts such as heterogeneity statistics and summary measures will be introduced. Interpretation and analysis of various graphical techniques will be discussed. Following this their power for detecting heterogeneity and outliers will be investigated thoroughly with simulated datasets.

Contents

1 Introduction **1**
- 1.1 History . . . 1
- 1.2 What is meta-analysis? . . . 2
- 1.3 Why conduct meta-analysis? . . . 4
- 1.4 Steps taken in meta-analysis . . . 4

2 Methods in meta-analysis **5**
- 2.1 Fixed and Random Effects . . . 5
- 2.2 Computing the Combined Effect . . . 7
- 2.3 Weighting of studies . . . 8
- 2.4 Presence of Bias in meta-analysis . . . 9
  - 2.4.1 Publication bias . . . 9
  - 2.4.2 Search and selection bias . . . 10
  - 2.4.3 Agenda Driven Bias . . . 10
- 2.5 Quantifying Heterogeneity . . . 11
- 2.6 Summary Measures . . . 12
- 2.7 Sensitivity analysis . . . 14

3 Graphical tools for meta-analysis **15**
- 3.1 Forest plot . . . 15
- 3.2 Funnel plot . . . 16
- 3.3 Linear regression . . . 17
- 3.4 L'Abbe plot . . . 19
- 3.5 Baujat plot . . . 20
- 3.6 Trim and fill . . . 21
- 3.7 Contour enhanced funnel plots . . . 23
- 3.8 Continuous data . . . 24

4 Simulation studies **26**
- 4.1 The Control Case . . . 26
- 4.2 Varying the number of trials . . . 27
- 4.3 Varying the size of the trials . . . 27
- 4.4 Adding publication bias . . . 29
  - 4.4.1 Significant case . . . 29
  - 4.4.2 Significant case, less bias . . . 30
  - 4.4.3 Adding publication bias (Insignificant case) . . . 32
- 4.5 Identifying an outlier . . . 33
- 4.6 Extension to continuous data . . . 35
  - 4.6.1 Continuous data with no bias added . . . 35
  - 4.6.2 Continuous data with publication bias added . . . 36
  - 4.6.3 Identifying an outlier . . . 37

5 Bayesian meta-analysis **40**
- 5.1 Bayesian meta-analysis . . . 40
- 5.2 Dissemination of failed reviews . . . 41
- 5.3 Non-Normally distributed meta-analysis . . . 41
- 5.4 Meta-regression . . . 41

6 Conclusion **43**

1 Introduction

Each layer of a wall contributes to its overall strength. It is safe to assume that each layer has a lower strength than all the layers combined. This is not the only example where this holds true, for example, imagine a randomised controlled trial where research was carried out to determine if a new drug has a desirable effect on a patient. More than one person would be included in the trial because the result from that individual couldn't be said to hold globally. The idea behind a clinical trial is to gather a random selection of people, which should represent the population, to obtain a result at the end. What if this idea can be made bigger? Imagine now the results taken from clinical trials are to be layers in a wall, the combination of these layers would make for an extremely strong wall indeed. This is the basic idea behind meta-analysis.

1.1 History

The origins of meta-analysis are a source of controversy due to disagreement as to what actually constitutes as the first real meta-analysis. The idea itself is not a new one, however without proper structure some would argue that the earlier attempts cannot truly be called meta-analysis. Either way it cannot be argued that these earlier attempts paved the way for the properly structured and statistically sound meta-analysis of today. The very first recorded attempt at combining studies can be traced back to Blaise Pascal in the 17th century. Pascal was a French mathematician whose main concern was applying maths to games of chance. A by-product of Pascal's research was a method for combining astronomical observations (NCBI).

It was an English mathematician, Karl Pearson, who would make the next significant step in advancing techniques in meta-analysis. Pearson was looking at incidence and mortality statistics in soldiers (who worked in both India and South Africa) for typhoid. Pearson, frustrated with the small groups that he had to work with and the poor statistical power the results would have, attempted to group the observations into larger series. The mathematics behind Pearson's combination of smaller studies would not be considered statistically sound by today's standards, however the results Pearson obtained are fairly similar to those which have been conceived through modern methods (Pearson 1904).

In the mid-20th century there were large quantities of data on extrasensory perception and with

this came a demand to combine these results into an overall effect size. In 1940 J.G. Pratt and J.B. Rhine published Extrasensory Perception Over Sixty Years, which combined studies spanning that time period. Unlike Pearson's work, the extrasensory perception studies were conceptually identical, which was very helpful from a statistical standpoint. This was the first recorded case where publication bias was accounted for and an attempt was made to correct it. However the methods used were not universally accepted with accusations of poor methodological design which led to misleading results.

The term meta-analysis was first used in a paper by Gene V. Glass in 1976, where it was defined as 'The analysis of analyses'. Glass was primarily interested in statistics within education and expressed concerns about the ever increasing amount of literature and information being wasted. Glass explained that traditional methods would no longer suffice without the aid of techniques for organising, depicting and interrelating data. Meta-analysis allowed Glass to answers questions single studies could not and to resolve disputes between studies where they arose. Glass concluded that it was embarrassing that researchers have less knowledge than that which has been proven, and further research must be carried out in the subject area of meta-analysis (Glass 1975). The problem of publication bias needed to be addressed and, among others, Robert Rosenthal was one to take charge. In 1979 he published a paper highlighting the problem. Rosenthal called it the *file drawer problem* where statistically positive results were more likely to be published than negative or insignificant results. Rosenthal decided that better bookkeeping was essential if meta-analysis was to progress and be respected as a statistically sound area of research (Rosenthal 1979).

A paper displaying the differences between random effects and fixed effects models was published in 1998 by Larry V Hedges and Jack L. Vevea. The two models were discussed in this paper and the conclusion reached was that the inference desired by the researcher should play a key role in deciding which model is chosen (Hedges 1998).

1.2 What is meta-analysis?

Many people conduct a version of meta-analysis without applying the mathematics aspect of it every day. If a student is given a piece of coursework requiring two factors to be weighed up, a common strategy would be to read into previous studies. The student could then summarise the findings and make judgements on which factors are of higher importance. Relationships between factors can

also yield surprising results. Subjective conclusions have effectively been drawn from the critical evaluation of reports and results available. However, with a lack of mathematical structure and consistency these conclusions may not represent the actual strength of these results or relationships. Meta-analysis was originally used in agriculture, education and vote counting, but without great impact. It was not until G.V. Glass introduced it to social sciences that it really took off. It's now a common tool implemented in genetics, drug testing and anything involving a clinical trial.

Formally, meta-analysis is a statistical technique used to combine results from several independent studies to get an overall effect size. In its most basic form it is a weighted sum of all the different effect sizes. How these weights are chosen depends on which model is used. In theory all studies are combinable if a researcher deems them so. Even different types of studies can be combined but if results are to be useful or meaningful, rigorous pre-set guidelines must be followed. There are two models that are going to be considered and the simplest of these is the fixed effects model. Here the effect size is assumed to be the same size across all the studies. The second model is called the random effects model and this assumes that the true effect size varies across the studies. However both these methods only hold true asymptotically.

Many people consider the biggest problem in meta-analysis to be heterogeneity. Heterogeneity is the measure of dissimilarity between studies, and of this there are two main types: methodological heterogeneity and simple heterogeneity. Both contribute to the overall statistical heterogeneity. Methodological heterogeneity comes from the naturally occurring diversity of the studies being combined. Studies will have different designs and assessment methods. Simple heterogeneity arises from differences in true effect sizes between studies and can be identified if there is more variation between the studies than that which is expected due to chance. If heterogeneity falls within the threshold of chance, researchers would often choose the simpler fixed effect model. However this is rarely the case. Dissimilarity between the studies is inevitable no matter how strict our methodical design. If significant heterogeneity is detected it is important that its source be found before moving forwards with a random effects model.

In meta-analysis there is usually two groups, the treatment group and the control group. The treatment group receive the treatment and their progress is recorded. The control group may receive a placebo or nothing at all and their progress is also recorded. It follows that the effect size is then the difference between the two groups in terms of progress. The effect size can take many forms, some of which will be explored later in this study.

1.3 Why conduct meta-analysis?

The reason we conduct meta-analysis is to improve the treatments we have available so as to find the best way of treating patients. Naturally, there is a lot of concern with the magnitude of the effect size and its direction (did it help or hinder). There are also regular disagreements between studies which a meta-analysis can settle and it can add consistency to the results. Type I and type II errors (see later) are a huge cause for concern in clinical trials, but with meta-analysis these are greatly reduced.

In smaller studies a larger treatment effect is needed for the effect size to be declared statistically significant (Sternea 2000). If a small study is carried out and there was an effect of a drug helping but it was a small effect size it would be very difficult to verify if it was actually helping or if this was simply due to chance. If trials are combined together and they all show a small effect size in the same direction then there is more reason to believe that this effect is due to the drug rather than being simply down to chance. This phenomenon is called statistical power and meta-analysis is a tool which can improve it. Statistical power refers to the likelihood of detecting, within a sample, an effect or relationship that exists within the population. More formally stated, 'the power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis, i.e., the probability that it will result in the conclusion that the phenomenon exists' (Cohen 1998).

1.4 Steps taken in meta-analysis

1. Form a clear focused review question.
2. Carry out an extensive search of all primary studies, if possible using a global language search.
3. Formulate a strategy to assess the quality of studies and check how they are

   relevant. Select studies and consider whether you will include unpublished studies.
4. Synthesis of study results and calculation of the effect size.
5. Interpret and analyse the results (Madhukar 2004).

2 Methods in meta-analysis

2.1 Fixed and Random Effects

When combining studies in meta-analysis there are two statistical techniques developed for inference. They are called the fixed effects and random effects models (there exists a third model called mixed effects model which we are not going to consider in this study) which have slightly different definitions and procedures, with guidelines available for choosing either one. Beyond affecting the computations, the models help define the goals of the analysis and interpretation of the statistics.

Fixed effect models assume that the effect size is the same across all the studies, the only variance here is the within study variance in each study. Weights are assigned according to the inverse of the variance so larger studies are more likely to take the lion's share of the weight. It is highly unlikely in real life for all studies to have a common effect size, although in rare cases datasets may be sufficiently close to this and the model will work well. *Random effects models* assume that the effect sizes are a random sample from the relative distribution of effects. Patients being older, more educated or from a wealthier economic background are likely to have different effect sizes giving way to two sources of variance, within study variance and between study variance. The weights are again assigned according to the inverse of the variance, except here the variance contains both between study and within study variance. This means that smaller studies are less likely to be trivialised. Heterogeneity and the statistical inference we desire will help decide which model is chosen.

Some researchers suggest heterogeneity is the principle factor in deciding which model to use, claiming that if heterogeneity is below a pre-set threshold then a fixed effects model should be chosen. Conversely if there is significant heterogeneity then a random effects model must be assumed (Hedges 1998). Although the two models give fairly similar results when there is a common population effect size, the inferences they give are different. Two types of inference will be considered for deciding a model, conditional and unconditional inference. If the researcher is mainly concerned with inference on the effect size parameter in the studies observed, we call this conditional inference. In this case a fixed effect model should be chosen. A potential drawback is that the inference gathered can only explain the studies included in the analysis and cannot be applied generally. Nothing can be said about past or future inferences. However the researcher may want to generalise the inference so that it holds globally using the selected studies. This assumes that the effect size is not a constant

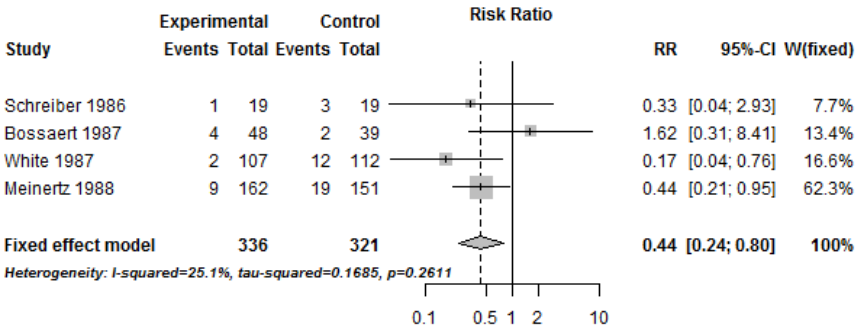

Figure 1: Example of fixed effect analysis

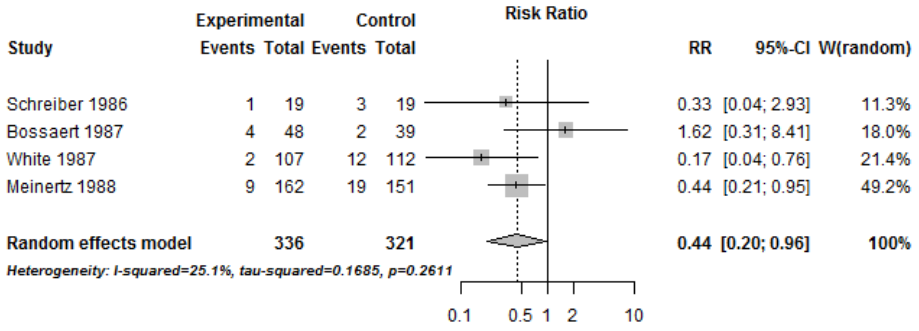

Figure 2: Example of random effects analysis

parameter shared by all studies and we call this unconditional inference. In this case the researcher should choose a random effects model.

Before continuing type I and type II errors should be recapped. A type I error is when a researcher's null hypothesis is true but it is incorrectly rejected. They are thought of as false positives. Type I errors are particularly dangerous, for example, imagine a new drug which is being tested for its ability to combat a nasty disease. A false positive will suggest that said drug does indeed combat the disease and should be manufactured for that purpose, but really it does not have the desired effect. Type II errors are when we fail to reject a null hypothesis which is false. This has less potentially harmful implications but is still misleading and must be minimised where possible.

Random effects models tend to have appropriate error rates for statistically significant tests at 5%, both when study sizes are the same (which is rare) and when study sizes vary. For fixed effects, where study sizes vary, Type I error rates are much higher than they should be and so confidence intervals based on these will be misleading (Hunter 2000).

Figure 1 and Figure 2 display the differences on a Forest plot (see later) between each model for the same data, Olkin95, for the treatment of myocardial infarction. After suffering a myocardial infarction, there are many treatments available, but one which is perhaps underused is thrombolytic

therapy. Thrombolytic therapy is the use of drugs to break up or dissolve blood clots and has been a major advance in the management of acute myocardial infarction. Unfortunately, it continues to be underused or is administered later than is optimal. The studies span from 1959 to 1990 and are mostly made up of small studies with 3 larger studies, and a few which are medium sized. From here on out this will be referred to as Olkin data and the desirable effect in the treatment group means the event occurs less frequently than in the control. Figure 1 displays the fixed effect model and Figure 2 displays the random effects model for the subset of studies 41,47,51,59.

The same data is used to highlight the differences between the two models. The random effects model has a higher confidence interval as expected but also notice the weight distribution on the very right hand column. Smaller studies gain more weight whilst larger studies have their weights reduced. This change of influence is one of the key differences between the models.

2.2 Computing the Combined Effect

Both random effects and fixed effects models use the inverse of the variance to weight each study, but how is this variation defined? There are several sources of variation throughout our analysis so to keep our analysis consistent the overall study variance will be used. Fixed effects models assume one true effect size meaning all studies have a common effect size, denoted by θ and the only source of variance is the within study error or estimation error, ϵ_i. The observed mean of each study is then represented as

$$Y_i = \theta + \epsilon_i \tag{1}$$

The overall study variance here is just the within study variance and is given by

$$V_i = \frac{\sigma^2}{n} \tag{2}$$

where σ^2 is the variance of the individual observations and n is the sample size. Clearly the variance decreases as n increases, i.e, if the sample size gets bigger we have a more precise estimate of the population parameter as a whole.

Random effects models assume that the studies are a random sample from the relative distribution

of effect sizes. So here μ is defined as our effect size instead of θ because it represents the mean of the distribution. There are two sources of variance here which will make up our overall study variance, the between study component and the within study component, denoted ζ_i and ϵ_i respectively ($\zeta_i = \theta_i - \mu$ where θ_i is the mean for study i and μ is the mean of all the studies). The observed mean Y for study i is then given by

$$Y_i = \mu + \zeta_i + \epsilon_i \tag{3}$$

2.3 Weighting of studies

The weight we assign each study is decided by the inverse of the variance so that more precise studies get more weight. We introduce τ^2 as the between study variance. The sample estimates of τ^2 are denoted T^2.

Fixed effects models only contain one source of variance, the within study variance. The weight for study i is then defined by

$$W_i = \frac{1}{V_i} \tag{4}$$

For random effects models we use the means of all the studies to compute the grand mean, μ, which introduces a second source of variance which we must include in our weighting of the studies. The between studies variance stays the same and is therefore a constant. The weight for study i is then defined by

$$W_i = \frac{1}{V_i + T^2} \tag{5}$$

where T^2 represents the between study variance. The within study variance continues to change from study to study. Note that if T^2 is zero then this is identical to our fixed effects model. It follows that our combined effect is found by

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \tag{6}$$

where M_F denotes fixed effects estimate and M_R denotes the random effects estimate. The variance

and standard deviations of M are found by

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} \tag{7}$$

and

$$SE_M = \sqrt{V_M} \tag{8}$$

Giving us 95% confidence intervals of

$$CI = M \pm 1.96 \times SE_M \tag{9}$$

(Borenstein 2009)(Hedges 1998).

The differences between fixed and random effects will now be shown using data (Else-Quest et al 2006) for the early development of gender differences. Using a fixed effects model, the combined effect returned was 0.0589 with a 95% confidence interval of (0.0242, 0.0935), and for the random effects model the combined effect returned was 0.0456 with a 95% confidence interval of (-0.0429, 0.1342). This again displays that although both models may have a similar mean, the random effects model will have a wider confidence interval.

2.4 Presence of Bias in meta-analysis

Meta-analysis has been subject to controversy from the outset. Some give it their highest praise whilst others question the validity of its techniques and results. One particular area of scrutiny is the potential sources of bias which appear at various stages of the analysis. Although it is impossible to eradicate bias completely, the results obtained should be as unbiased as is statistically possible. With less bias comes less heterogeneity, more precise results and better meta-analyses.

2.4.1 Publication bias

Publication bias is generally considered to be the most common and often the most difficult to eradicate. Ideally if a researcher carries out an analysis the results should be published. Then

picking studies to combine in a meta-analysis would be a simple lottery. This however is not the case. Studies which are statistically positive/significant are more likely to be published, published quickly, published in English, cited by others and published in well-known journals. The main concern is that particular studies never get published. This causes a shift in the effect size towards positive or desirable findings which results in misleading meta-analysis. This situation can be dangerous. Imagine a treatment which has no effect, studies showing a positive result are being published whilst some of the negative studies are not. A meta-analysis of the published studies would yield a positive result and unhelpful treatments may be administered (BMJ 1998). Publication bias can be identified with relative ease if one thinks about the process of burying an undesirable study. Large studies with thousands of patients are likely to be in the public eye, so burying them would pose a difficult challenge. Conversely smaller studies don't have such a high profile and so can disappear relatively easily. This leads to a very interesting observation, that smaller studies are going to have a higher rate of publication bias, and statistical methods for spotting this will be discussed later.

2.4.2 Search and selection bias

To conduct meta-analysis a researcher must first find studies which are going to be included. This introduces a potential source of bias. Meta-analysis published in English are often exclusively based on trials published in English (BMJ 1998). There is no reason for not including trials published in different languages. To avoid this a 'Global language search' can be used. Armed with studies from a global language search the researcher must then pick the inclusion criteria which can potentially lead to another source of bias. If the criteria are picked by an expert with knowledge in the area, the criteria may then be influenced by knowledge of previous trials. The inclusion criteria must not be manipulated with the intention of increasing the probability of including a study which yields a significant result.

2.4.3 Agenda Driven Bias

The main objective of most corporations is to make money. So if a business finances a clinical trial or provides a trial with their products, the studies may be affected by bias. This could involve manipulating the data or not publishing negative results which may harm the sales of that product or service. If a meta-analysis is carried out, sponsors should be named so the credibility of the results

can be checked. Corporations behind the analysis cannot be allowed to interfere. The World Health Organization (WHO) has created the International Clinical Trials Registry Platform (ICTRP) to help combat bias. If signed up to the register then details of the methodological procedure must be uploaded and the results of the study must always be published. This stops researchers manipulating procedures to get a significant result and all studies are published. Their sponsors must also be named along with other regulations to stop misleading meta-analyses. Unfortunately it is not mandatory to register, but fortunately this is an increasing trend. This step is in the right direction for improving the credibility of meta-analysis.

2.5 Quantifying Heterogeneity

The difficulty heterogeneity poses to meta-analysis has already been discussed and to statistically assess the difficulty we need some way to quantify it. Two types of heterogeneity (simple and methodological) are combined to give an overall statistical heterogeneity. Random effects meta-analysis can be used to give some measure of heterogeneity because it accounts for between study variance, and so a simple estimate could be found from an odds ratio or risk ratio. However this estimate is rather poor and does not allow for any comparisons to be made, such as that between binomial and continuous data. A common method for quantifying heterogeneity is Cochrane's Q-test, where a p-value is given as an indication of the extent of between study variability. However this test has a low power for detecting heterogeneity when there are too few studies and superfluous power to detect clinically unimportant heterogeneity. The problematic nature of this test means it is usually not taken as a relevant summary when quantifying heterogeneity, but it can be used to derive more appropriate measures. Here the precision is assumed to be known, however in real life we get this from the data. A homogeneity test for θ_i is given by

$$Q = \sum w_i (y_i - M_F)^2 \tag{10}$$

with $k-1$ degrees of freedom on a χ^2 distribution, where k is the number of studies. Q is poor at detecting true heterogeneity. The expectation of this is:

$$E[Q] = T^2(\sum w_i - \frac{\sum w_i^2}{\sum w_i}) + k - 1 \tag{11}$$

Which in the absence of heterogeneity gives $E[Q] = k - 1$

A better measure for heterogeneity is H^2 which is given by

$$H^2 = \frac{Q}{k-1} \tag{12}$$

H^2 is better because it describes the relative excess of Q over its degrees of freedom. Higher levels of H^2 mean higher levels of heterogeneity. If $H = 1$, this suggests homogeneity since $E[Q] = k - 1$ (Higgins 2002). An even better statistic to measure heterogeneity is I^2

$$I^2 = \frac{H^2 - 1}{H^2} \tag{13}$$

The I^2 statistic describes the percentage of variance that is not down to chance, but rather heterogeneity. I^2 provides a measure of the consistency between studies and is usually expressed as a percentage. Higher values of I^2 mean higher levels of heterogeneity.

At least one of these statistics should always be present in meta-analyses.

2.6 Summary Measures

The selection of a summary measure for binary data has always been an area of controversy. There are a range of summary measures which can be used in a meta-analysis. The risk ratio or relative risk (RR), odds ratio (OR) and risk difference (RD) are all going to be considered here. RR and OR are relative and RD is absolute. Each summary statistic corresponds to a different pattern of predicted absolute benefit of treatment with variation in baseline risk. The factors considered in making this decision are:

1. Consistency of effect
2. Ease of interpretation
3. Mathematical properties

(Deeks 2002). A problem arises because the summary statistic which has the best mathematical properties (OR) is difficult to understand and interpret. Many meta-analyses have been carried out assuming RR has been used in analyses when in fact OR has been. The problem here arises when

researchers are keen to get people interested in undertaking research and so intuitive understanding of the summary measure is very useful. Here RR would be suggested. Conversely other researchers claim that better mathematical properties are of paramount importance and ease of interpretation is of little value, and so suggest OR. RD is actually the easiest to interpret but unlike the other two it isn't relative and so lacks consistency. For this reason it is usually not used in meta-analysis as absolute values are unlikely to be generalisable. There is very little difference between OR and RR and so deciding between them can be a difficult choice. A possible solution would be to try both out and choose which ever has the smallest heterogeneity statistic, but this doesn't work in cases where there are too few studies. For ease of interpretation and consistency RR is the best compromise and will be used throughout this study unless otherwise stated.

	Event 'success'	No Event 'Fail'	
Treatment Group	S_E	F_E	N_E
Control Group	S_C	F_C	N_C

Table 1: Clinical trials results

The results from clinical trials can be captured such as in Table 1

From here, the summary statistics can be calculated

$$RR = \frac{\textit{Risk of event in experimental group}}{\textit{Risk of event in control group}} = \frac{S_E/N_E}{S_C/N_C} \tag{14}$$

$$OR = \frac{\textit{Odds of event in experimental group}}{\textit{Odds of event in control group}} = \frac{S_E/F_E}{S_C/F_C} \tag{15}$$

$$RD = \textit{Risk of event in experimental group - risk of event in control group} = \frac{S_E}{N_E} - \frac{S_C}{N_C} \tag{16}$$

Figure 3 displays the three different summary measures used in a funnel plot.

Both OR and RR have a value for I^2 at 69.9% with respective confidence intervals [46.8% ; 83%] and [46.3% ; 82.8%]. This shows how similar the two summary measures are in terms of heterogeneity. The value of I^2 for RD was 76.6%, considerably higher than the relative summary measures, but its

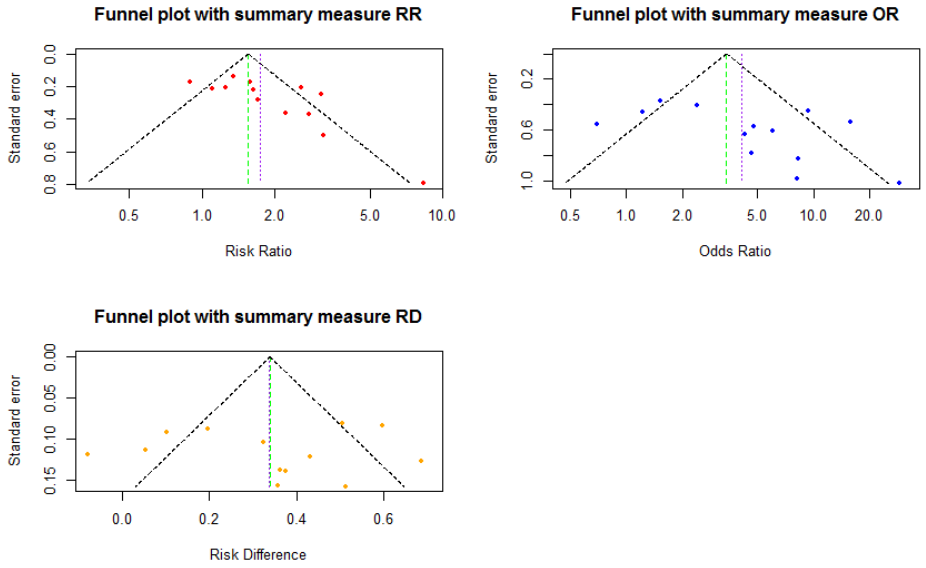

Figure 3: Cisapride data plotted with different summary measures

confidence interval was much smaller, [60.1% ; 86.3%]. Notice also that the effect size estimate is much higher for OR than RR.

2.7 Sensitivity analysis

A term very often found in conjunction with meta-analyses is 'sensitivity analysis'. This is a cautious approach to 'correct' for publication bias in which we draw conclusions from meta-analyses under different study groups (not to be confused with sub group analysis), treatments or patients and look at plausible possibilities for the extent of publication bias. These are then compared with each other and with standard methods for obtaining results. The aim is to embed a standard model with a range of different models which are plausible in the context of the data, and the range of these different models are then compared with the original model (Copas 2000).

3 Graphical tools for meta-analysis

Bias and heterogeneity are present in all meta-analyses of clinical trials. Sometimes it's a symptom of the clinical trials which has manifested itself in the meta-analysis, and sometimes it's a problem with the meta-analysis itself. Bias must be minimised and eradicated completely where possible. In order for this to be done methods are needed to spot bias, remove it, and adjust for it before re-estimation of the effect size. All summary measures in this study will be taken as risk ratios unless otherwise stated and datasets are binomial unless otherwise stated.

3.1 Forest plot

These are sometimes also known as confidence interval plots. Forest plots are perhaps the best known graphical tool in meta-analysis and they show effect estimates and the corresponding confidence intervals for each study. A forest plot is drawn in Figure 4. The middle of each green square on the right hand side of the plot is the effect estimate, and the larger this square the more weight has been assigned to it. The line through the block corresponds its 95% confidence interval. The overall effect is represented by a red diamond with its confidence interval being the width of the diamond. To the right of this is the numerical representation of the effect size and the confidence interval. The purple bar below this represents the size of the prediction interval, and weights assigned to each study are on the very right hand side of the graph (Anzures-Cabrera 2010). Some heterogeneity statistics are printed on the bottom left for additional information. The line in the middle of the Risk Ratio section divides 'favours control' with 'favours treatment'.

The data used was collected in 2001 and was in relation to cisapride (Hartung 2001). Cisapride, a gastrointestinal tract promotility agent, can cause life-threatening cardiac arrhythmias in patients susceptible either because of concurrent use of medications that interfere with cisapride metabolism or prolong the QT interval, or because of the presence of other diseases that predispose to such arrhythmias (Smalley 2000). From here on out this will be referred to as cisapride data. The desirable treatment effect is increasing the number of events in the treatment group.

The data has been ordered smallest effect size to largest to help with analysis. It can be seen that the fixed effects model has a much smaller confidence interval and with the weighting primarily assigned to larger studies. Also notice that the larger studies tend to have a treatment effect closer to the

middle.

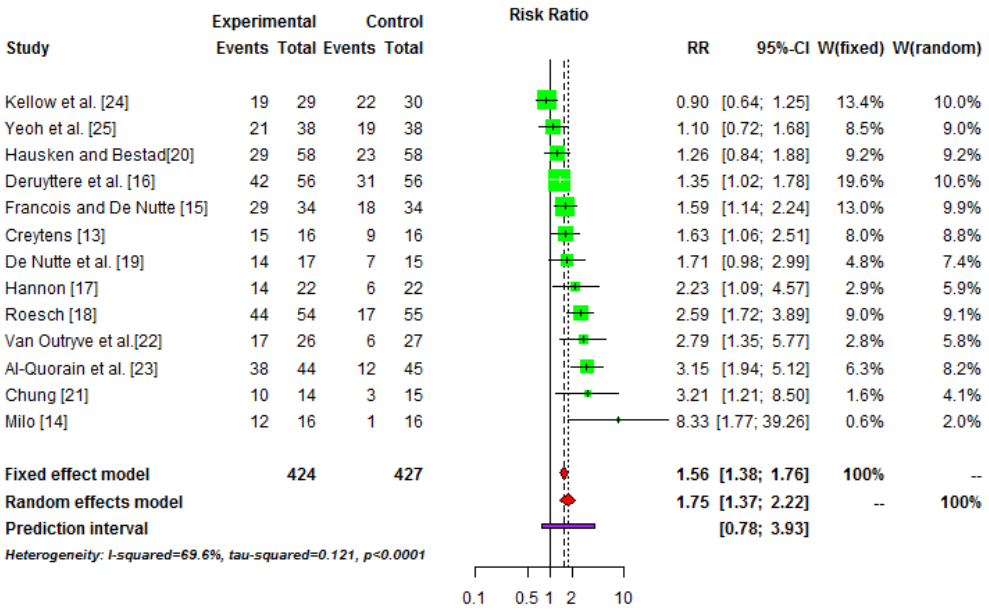

Figure 4: Forest plot for cisapride data

3.2 Funnel plot

Meta-analysis has the advantage of using many studies to get an overall treatment effect. This process should increase the precision of the effect size due to the larger number of patients considered. However it has been the case that clinical trials with a large number of patients often contradict the results in a meta-analysis. The result isn't as shocking as it seems when bias is considered. There are many types of bias trials may succumb to, with publication bias being the most frequent offender. Larger studies have received a greater investment of money and time which means they are more likely to be of high methodological design and published regardless of the result. Conversely there is a tendency for smaller studies to be more prone to selective suppression and less thought put into their design (Stern 2000).

Funnel plots plot the inverse of the standard error against an effect estimate (Egger 1997), in this case RR. Its name arises from the fact it looks like an inverted funnel, with smaller studies scattered widely at the bottom and larger studies becoming narrower as standard error decreases. This decrease in scatter is due to increased precision of the clinical trial. In the absence of bias the funnel plot will be symmetric, but in the presence of bias it will skew to one side or another and be asymmetric.

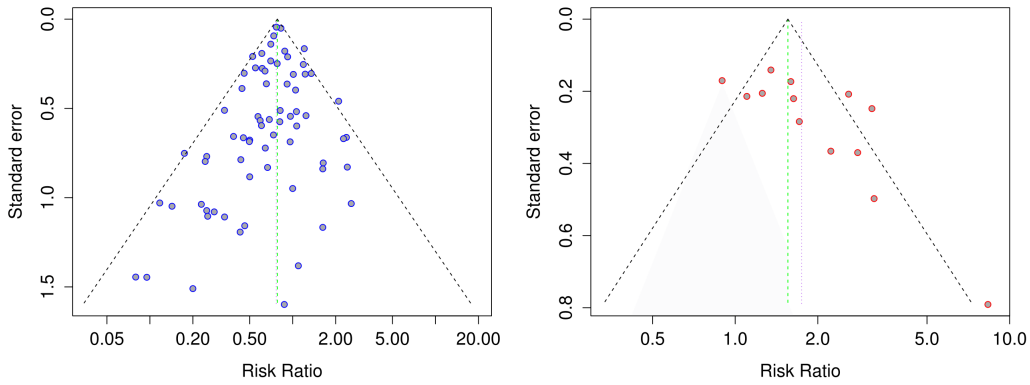

Figure 5: Left:Funnel plot for the Olkin data; Right: Funnel plot for the cisapride data

For example, if publication bias is present we expect there to be a skew in the direction of the desirable finding where negative or insignificant findings may not be published. A funnel plot should not be used for any kind of analysis if there are less than ten studies (Stern 2011). The value of analysing a funnel plot is not exact but it can give a good indication of any potential issues within the meta-analysis.

Figure 5 displays the Olkin data and cisapride data respectively. The Olkin data is an example where there is very little bias present, however there appears to be a small area in the bottom right corner with no studies but this isn't extreme. On the other hand, the cisapride has a lot of its smaller studies on the right hand side with none on the left hand side. However the larger studies are fairly evenly spread out. From inspection it seems that there is publication bias present in the cisapride data. The random effects estimate is a purple dotted line whereas the fixed effect estimate is in green.

The problem with Funnel plots is that they give no statistical evidence of heterogeneity or bias and so we must look to other methods for analysing this in a quantifiable manner. Linear regression tests may have the answer.

3.3 Linear regression

Linear regression methods are used to quantify the bias in meta-analysis which corresponds to analysis of a Galbraith plot or radial plot, although the regression does not necessarily run through the origin. Egger suggested a method which regresses the standard normal deviate (or Z score, where the standard normal deviate is $x + y\times$ precision) against precision (Sterne 2000). Smaller studies

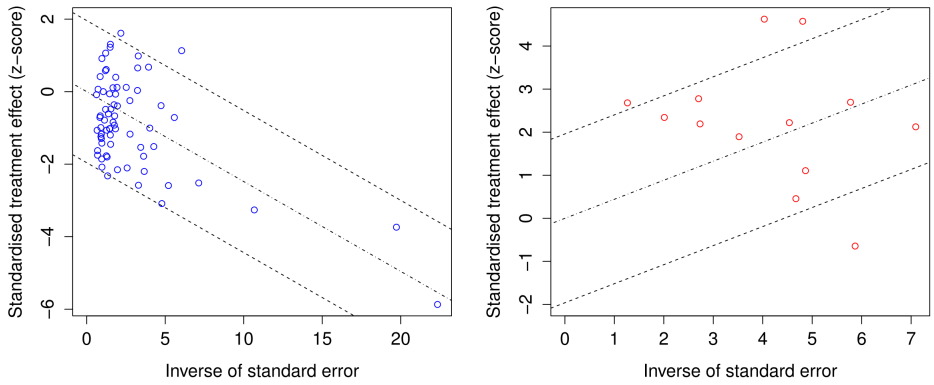

Figure 6: Left: Galbraith plot for the Olkin data; Right: Galbraith plot for the cisapride data

have lower precision and a smaller standard normal deviate (SND) and so will be close to the origin. Conversely larger studies will be more precise and will have larger SNDs if the treatment is effective (Egger 1997). Thus, if unaffected by bias, the regression will pass through the SND zero (i.e. the origin) with the slope of y indicating the magnitude and direction of the effect size. Conversely if bias is present, it will not pass through the SND zero (i.e. not through the origin) and the distance it is from zero is indicative of how much bias there is. If it passes through the y axis at $y > 0$ this means that bias is affecting the effect size in a negative way, and if $y < 0$ this means it's affecting the effect size in a positive way. To aid our interpretation of the analysis lines are added which represent the 95% confidence interval, if more than 5% of the studies are outside this threshold it is a sign of heterogeneity. The radial plot also displays the direction of the treatment effect, and an upward slope represents an increased event frequency, whereas a downward slope represents a reduction in event frequency. The Radial plot for both the Olkin and cisapride data are plotted in Figure 6.

Both sets of studies seem to go through zero which is a good sign, this could be an indicator that there is little heterogeneity. However with the cisapride data, notice that 6 of the 13 studies ($> 5\%$) are outside the 95% confidence interval. This is a strong indication of heterogeniety. Conversely with the Olkin data only 3 of 70 studies are outside this range, which is around 5% as required. The regression line for the Olkin data slopes downward suggesting a reduction in event frequency for the treatment group, which is the desirable effect. Conversely for the cisapride data the regression line slopes upward, an increase in event frequency for the treatment group, which is again the desirable result. This suggests that both treatments are having a more desirable effect than their respective control groups.

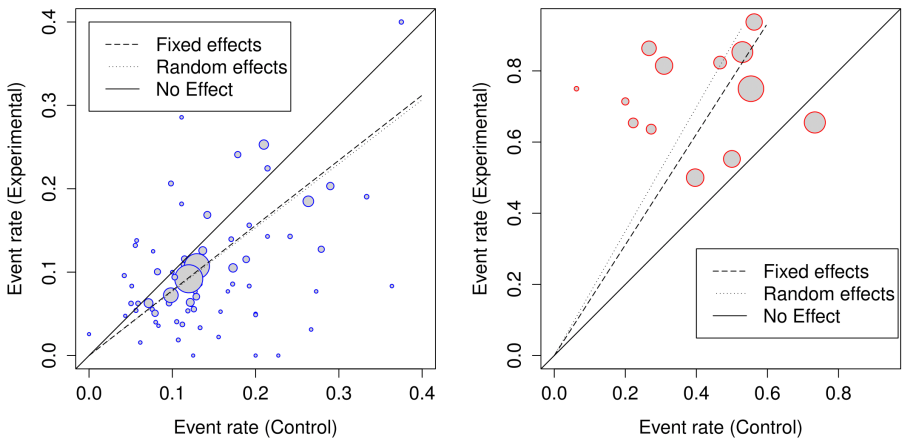

Figure 7: Left: L'Abbe plot for the Olkin data; Right: L'Abbe plot for the cisapride data

3.4 L'Abbe plot

L'Abbe plots can only be used on a binomial dataset with risk ratios and were first seen in 1987. It plots the observed risks for both control group on the x axis and the treatment group on the y axis and is used to detect heterogeneity. A regression line may be plotted through the graph which indicates no effect. The dots are now sized in proportion to the study weights as there is no measure of variance or precision anywhere else on the plot. This allows us to inspect the range of risks among the studies and to highlight heterogeneity. Clustering of control group risks along the x axis suggest that our assumptions about baseline risk are broken and so heterogeneity is present (Anzures-Cabrera 2010). The plot can give an indication on which summary measure to use to get more consistent results. Outliers can be identified by their distance away from the effect estimate lines. One suggestion here would be to remove these outliers one by one until heterogeneity is no longer significant (Song 1999).

Figure 7 displays L'Abbe plots for the Olkin data and for the cisapride data.

For the Olkin data we see that our fixed effect estimate and random effects estimate are both below the line of no effect suggesting that events were happening less frequently in the treatment groups than in the control group. Conversely with cisapride data only one study is below the line of no effect whilst the rest are above, suggesting more treatment group events are occurring. There appears to be one small outlier for the cisapride data at the top left. There is no strong suggestion of heterogeneity for either study here but Baujat plots can be useful in analysis.

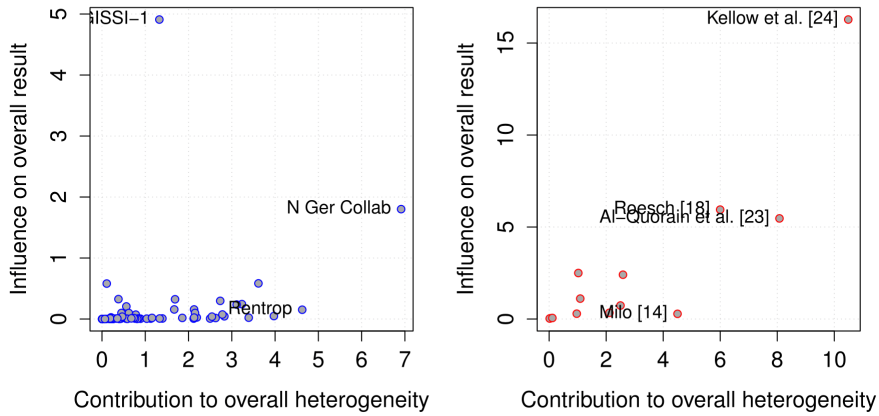

Figure 8: Left: Baujat plot for the Olkin data; Right: Baujat plot for the cisapride data

3.5 Baujat plot

Another graphical method to check for heterogeneity is a Baujat plot. First introduced in 2002 it was well received by most statisticians. It is very effective for finding which studies are the source of heterogeneity in meta-analysis. On the x axis is the contribution to the overall heterogeneity statistic (Cochranes Q statistic), and on the y axis the standardised difference of the overall treatment effect.

A plot is drawn with and without each study to measure the contribution to the overall effect size of each study (Baujat 2002). Further to the right and top indicates high heterogeneity and influence of result. A Baujat plot is important when a researcher wants to bypass subgroup analyses and immediately find which study is causing the heterogeneity. The Baujat plots for both sets of data are plotted in Figure 8.

To aid graph visual analysis, only study names with extremes of both heterogeneity and influence of effect size are printed. With the Olkin data it appears there is only one result with extremely high influence on the overall result *(GISSI-1)*, with *White* being the only study affecting heterogeneity to a high degree without also affecting influence on the overall result in a similar way. The rest of the studies look normal. With the cisapride there is also one study causing heterogeneity without affecting influence on overall result to a similar degree, *Milo [14]*, but not enough to be considered an outlier. The rest of the results do not appear to be cause for concern.

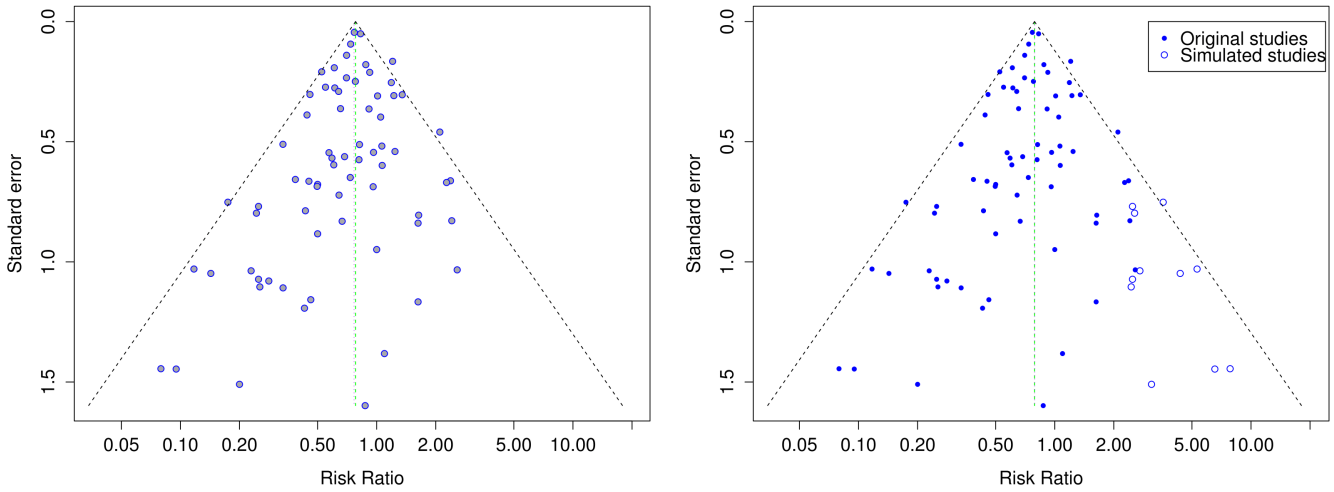

Figure 9: Trim and Fill plot for the Olkin data

3.6 Trim and fill

A difficulty in meta-analysis is that we must collect all studies to do with that meta-analysis, published and not published, if inferences are to be valid (Egger 1997, Duval 2000). A representative sample must be taken for credible meta-analysis to be carried out, and to get a representative sample we must have a representative population from which to choose a sample. When publication bias is detected a researcher should find the source of it and try to remove it. If this can't be done other methods are needed to correct for high levels of heterogeneity.

A nonparametric, graphical technique available is called trim and fill. If a funnel plot is undergoing a trim and fill, first the asymmetric funnel plot has a line drawn up the centre. Then smaller studies which are making the funnel plot asymmetric are removed temporarily and the new, 'true', centre of the graph is found and drawn. The studies which had been removed along with their mirror images are replotted (the mirror being the new centre of the graph not the old centre) and analysis is done on the original studies and these added studies. The added studies represent the studies which are assumed missing due to publication bias.

This method heavily relies on the assumption that there should be a symmetric funnel plot, and there is no way to tell if the adjusted effect size would match that of the meta-analysis had there been no publication bias, giving this method a low power. There are other reasons why the funnel plot is asymmetric other than publication bias which this method neglects. This method has also been proven to be very misleading when there are high levels of between study variance (Peters 2007). Trim and fill for both sets of data are shown along with their original funnel plots for comparison in

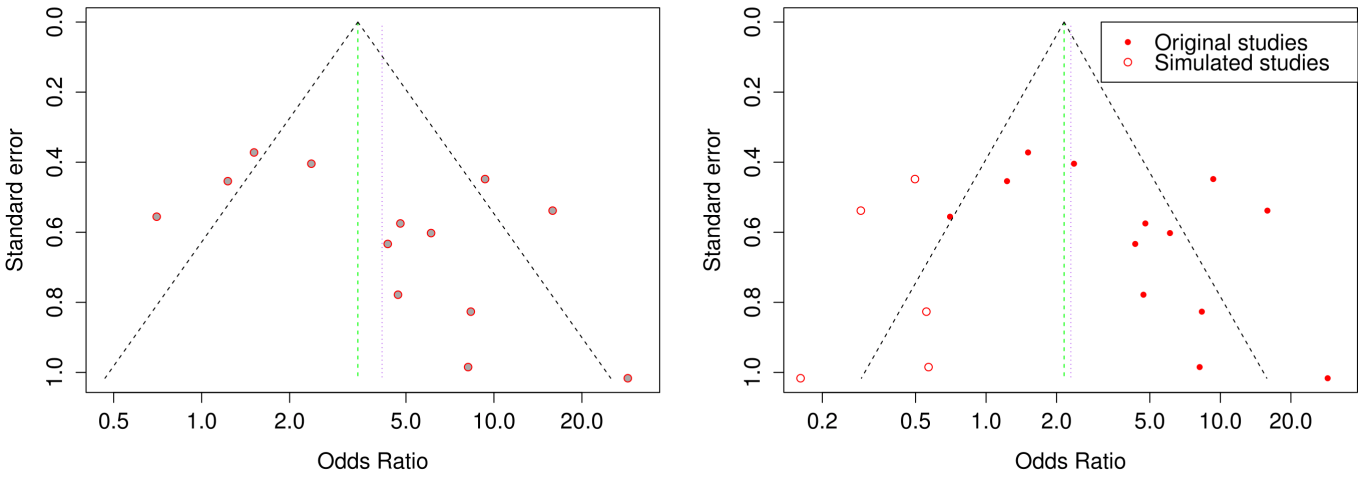

Figure 10: Trim and Fill plot for the cisapride data

Figure 9 and 10.

From inspection it appears that the random effects and fixed effects estimates have been brought closer together in Figure 3.6. The graph looks more symmetric but these estimates aren't that much different to before, suggesting that trim and fill was not necessary here. This was not unexpected as there were already low levels of heterogeneity with this data set. The data before addition of simulated studies had a value of $I^2 = 18.6\%$ with a 95% confidence interval of [0% ; 40.1%]. With the 10 simulated studies added $I^2 = 26.6\%$ with a 95% confidence interval of [2.6% ; 44.7%]. Trim and fill has increased the amount of heterogeneity. The random effects estimate from the original funnel plot was 0.7323 and this increased to 0.7565 which is expected, as all the studies added into the plot were above the original estimate. The 95% confidence interval also increased from [0.6638 , 0.8078] to [0.6785 , 0.8434]. The trim and fill technique has not been beneficial to the analysis in this case.

Figure 10 shows the cisapride data and again we see that the estimates for both random effects and fixed effects are closer. This funnel plot looks much more symmetrical but some studies are now quite far away from being included inside the 95% region. The heterogeneity here was quite high to start off with, at 69.9%, and this increased to 79.9%. The respective confidence intervals were [46.8% , 83%] and [69% , 87%] so there was a decrease in this, which is perhaps the only positive thing from this trim and fill. A researcher must use discretion when deciding whether to implement this method.

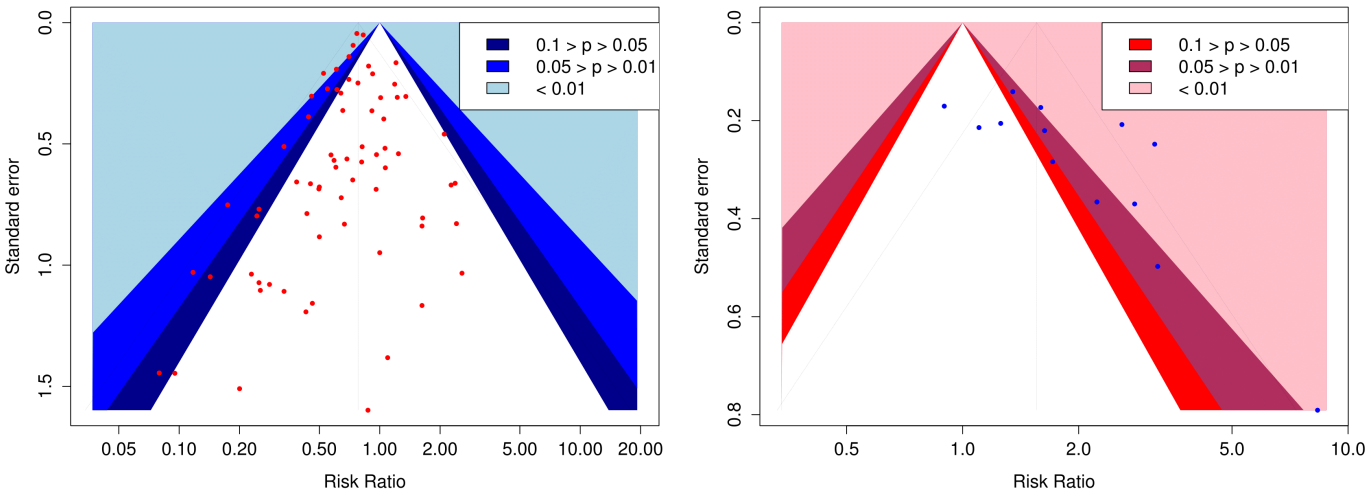

Figure 11: Left: Contour enhanced funnel plot for the Olkin data; Right: Contour enhanced funnel plot for the cisapride data

3.7 Contour enhanced funnel plots

Deciphering which type of bias that is present in a particular analysis can be very difficult, but contour enhanced funnel plots can aid in doing just that. The mechanism for how publication bias works is important here. Statistically nonsignificant papers and statistically negative papers are less likely to be published. Of the two, more often it is statistically nonsignificant papers that are not published rather than negative findings (Peters 2008). Outcome reporting bias also plays a part. Presence of outcome reporting bias suggests that even in papers which are published, only the significant outcomes are reported.

When funnel plot asymmetry is found, there may be reasons other than publication bias that are responsible. Any factor which affects study variance or effect size could act as a confounding agent. Contour enhanced funnel plots highlight the different areas of statistical significance. If the asymmetry is in a place of low significance it is more likely that the asymmetry is due to publication bias, but if it is in a place of high significance, bias is more likely to do with poor methodological design. Another case is where there is a 'tunnel' effect, where there is an absence of studies up the middle of the funnel plot caused by suppression of nonsignificant papers. This type of suppression will leave a higher number of studies sitting on or outside the significance lines.

The contour enhanced funnel plots are plotted for the two sets of data in Figure 11. Blue dots with red contours represent the Olkin data and red dots with blue contours represents the cisapride data. The Olkin does not seem to have much asymmetry. If any exists it's in a nonsignificant zone

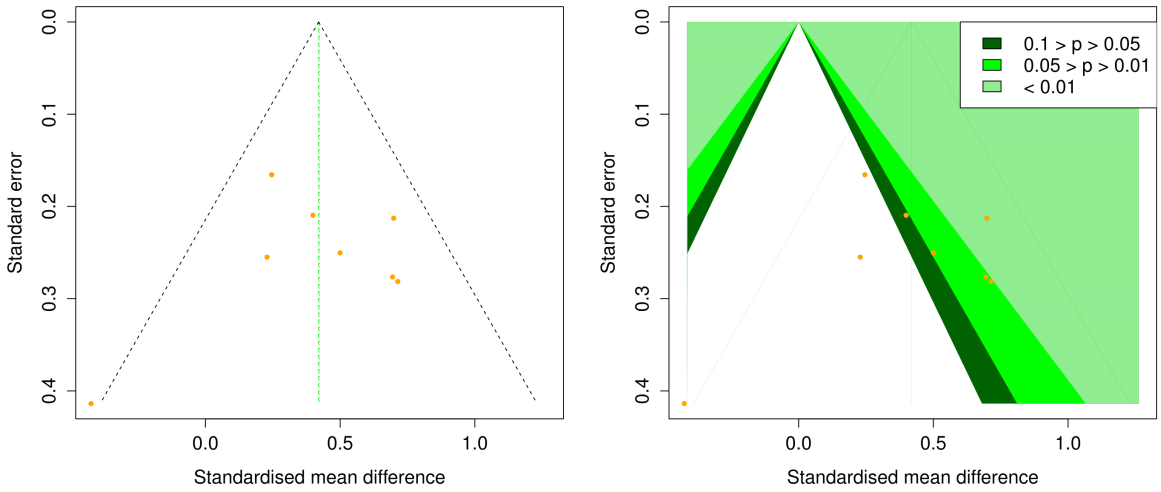

Figure 12: Continuous data funnel plots for amlodipine

on the right hand side of the graph which indicates possible publication bias. The cisapride data shows clear asymmetry, with missing studies within the nonsignificant zone indicating the presence of publication bias.

3.8 Continuous data

The extension to continuous datasets will be explored in this section. The summary measures for continuous data are the mean difference (MD) and standardised mean difference (SMD). In this study SMD will be used because standardised summary measures can be applied more generally than absolute summary measures. MD is just

$$MD = mean\ of\ treatment\ group - mean\ of\ control\ group \tag{17}$$

However, the standardised mean difference is

$$SMD = \frac{MD}{pooled\ standard\ deviation\ of\ both\ groups} \tag{18}$$

Plots drawn with continuous data are drawn in the same way as for those with binary data except with different summary measures. There are some plots which cannot be drawn with continuous data such as a L'abbe plot.

First we need to introduce a continuous dataset. This data set will be referred to from now on as the

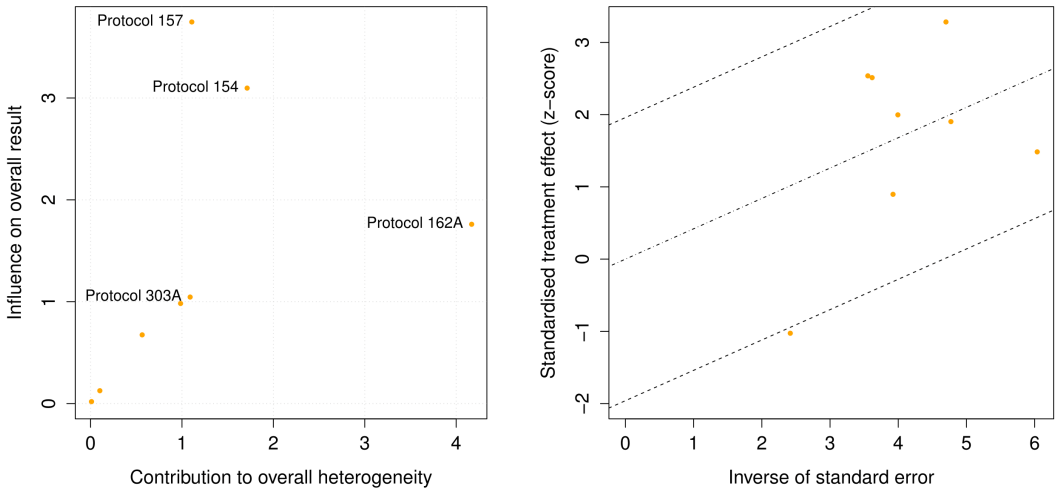

Figure 13: Left: Baujat plot for amlodipine; Right: Galbraith plot for amlodipine

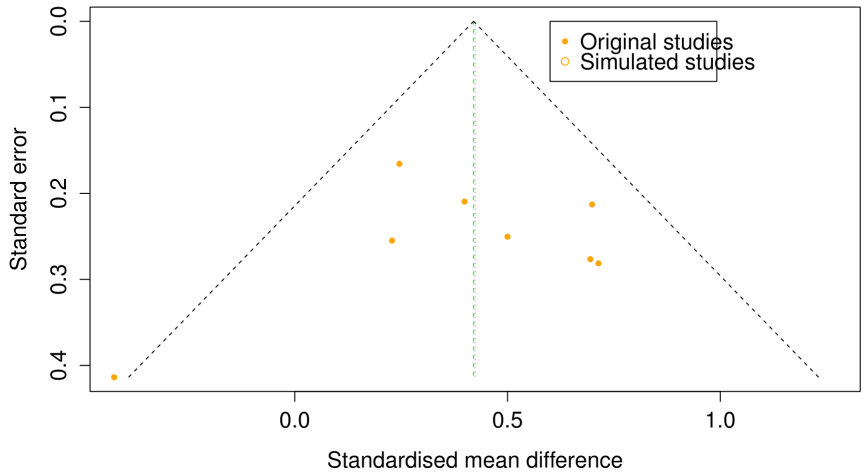

Figure 14: Trim and Fill plot for amlodipine

amlodipine dataset. Amlodipine is in a group of drugs called calcium channel blockers. Amlodipine relaxes blood vessels and improves blood flow. It is used to treat hypertension or angina among other conditions. The data was taken from (Hartung 2001). Although the data set is too small to be used for funnel plot analysis, the following plots in were generated from the amlodipine dataset to display the similarity of the methods. Figure 12 shows the funnel plots. The Baujat plot and Galbraith plot are displayed in Figure 13.And finally Figure 14 displays the trim and fill funnel plot for continuous data.

4 Simulation studies

Simulating studies with different parameters will give examples of how the graphical techniques from section 3 can detect different sources of heterogeneity and bias. In this section, different pseudo datasets will be compared using whichever graphical techniques are appropriate. The probability of event for treatment and control groups were taken to be 0.6 and 0.4 respectively for the binary data simulations. This means a beneficial treatment effect is more events occurring in the treatment group. Each data set will have 20 trials with 1000 overall patients in the meta-analyses with the exception of section 4.2 which varies the number of trials to measure this effect. The figures stated will be found using a for loop with 10,000 iterations for consistency and precision.

4.1 The Control Case

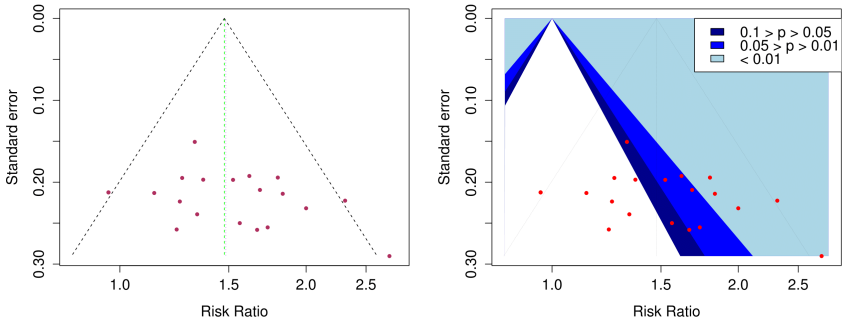

Figure 15: Funnel plots for control case

This example is very basic and will be used for comparisons. There are 20 trials with 50 in each trial, and all the trials are the same size so any heterogeneity will simply be due to chance. Here I^2 is expected to be small or zero because there is nothing to cause heterogeneity and for the funnel plots to be asymmetric. These are displayed in Figure 15

There does not seem to be any suggestion of asymmetry with the studies fanned out as expected. The heterogeneity statistics are: I^2 =0%, Q= 13.91 which is consistent with funnel plots. The 95% confidence interval for I^2 is [0%, 28.9%] and the p-value for heterogeneity comes in at 0.7887, which is non-significant. The null hypothesis that there is no heterogeneity cannot be disproved and so we infer no indication of heterogeneity. The fixed effects and random effects estimates are the same showing very little between study variance. No further analysis is needed here.

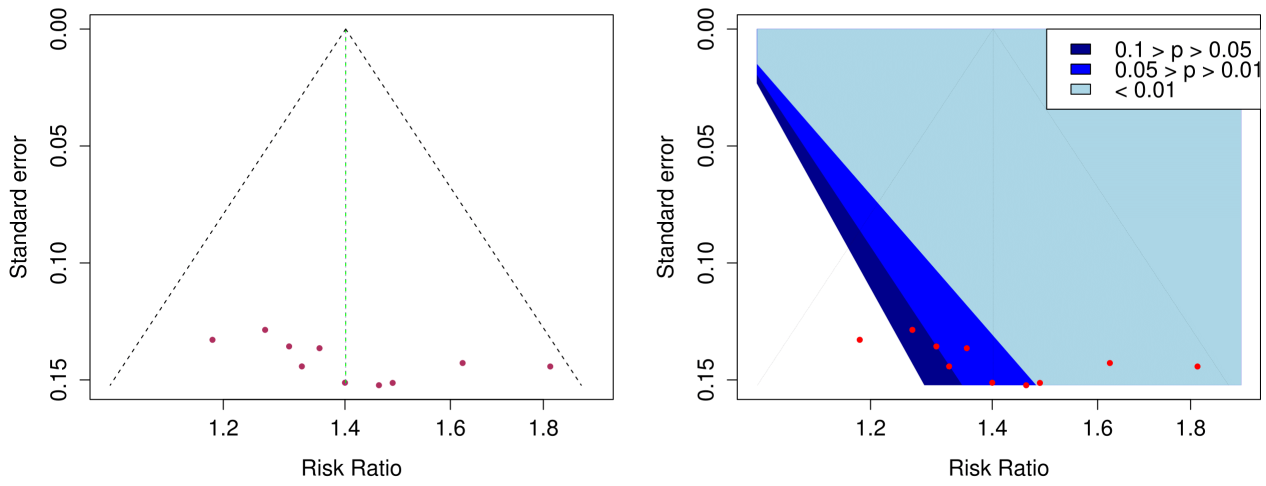

Figure 16: Funnel plots with fewer studies

4.2 Varying the number of trials

Here the number of trials was decreased to 10 and the number of patients in each trial increased to 100 so that the total number of patients remain at 1000. The heterogeneity statistics are expected to be around the same value, maybe a little higher with larger confidence intervals because one trial which leans to one extreme will have more influence on the analysis. The funnel plots are displayed in Figure 16.

There does not appear to be any asymmetry here even though there are 6 studies on the left hand side and only 4 on the right hand side of the funnel plot. This is because the studies on the left hand side of the effect size estimate are closer to this estimate. Conversely those on the right are more extreme values. The heterogeneity statistic as expected is around the same value as before but with a bigger confidence interval, 0% and [0% , 52.1%]. The value for Q came in at 10.17, which is considerably smaller because there are fewer studies and so fewer degrees of freedom. The p-value for heterogeneity came in at 0.632, again our null hypothesis cannot be disproved. No further analysis is needed here.

4.3 Varying the size of the trials

Instead of all the trials being the same size it was decided to vary the size of the trials. There are 10 trials with 30 patients, 6 trials with 60 patients, 2 trials with 75 patients and 2 trials with 95 patients. Having different sized trials could increase the amount of dissimilarity between the trials

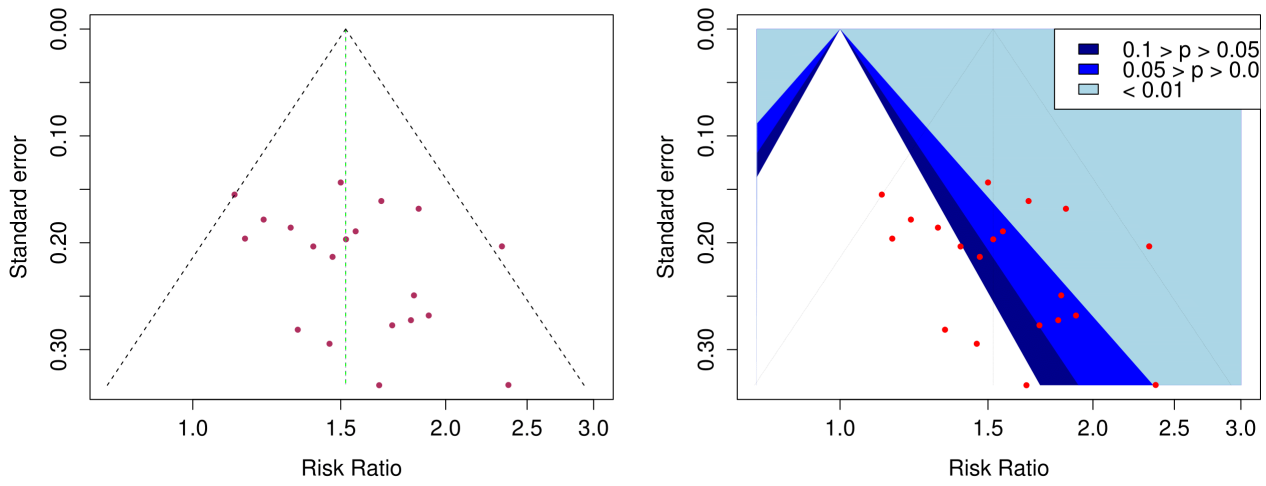

Figure 17: Funnel plots with different sized studies

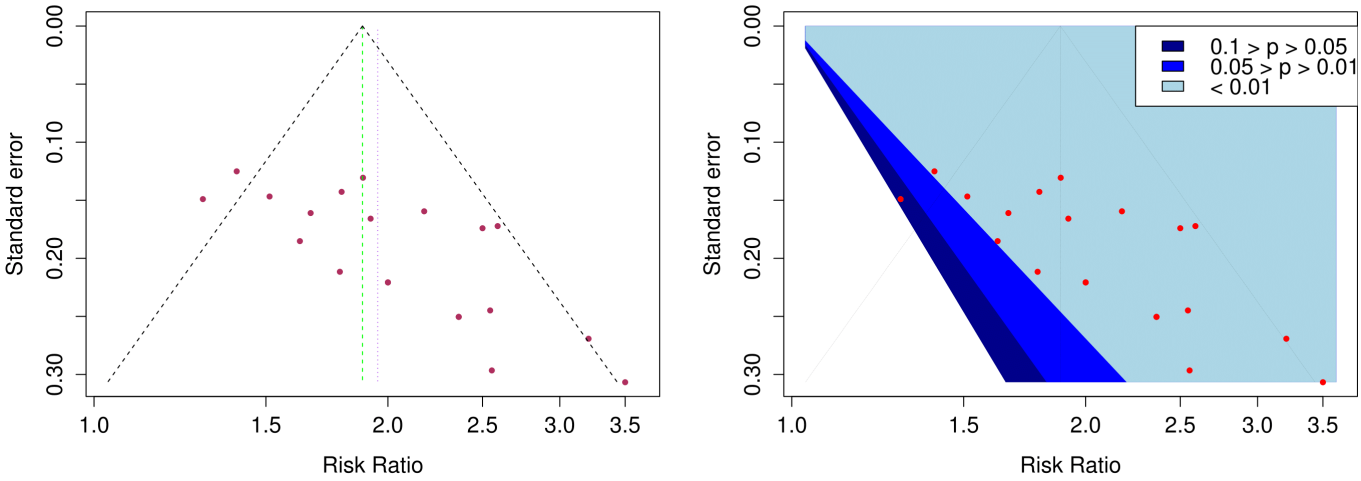

Figure 18: Funnel plots with added publication bias

and so increases heterogeneity. The funnel plots are displayed in Figure 17.

The value for I^2 came in at 0.2% with a confidence interval of [0% , 48.1%]. The value for Q was 19.05. The p-value came out at 0.4539, so still there is not significant enough heterogeneity to cause concern. The fixed and random effects estimates were very close again. From this it is clear that difference in trial size has little effect on heterogeneity, so if bias is not present, significicant heterogeneity is unlikely as any disimilarity is due to chance.

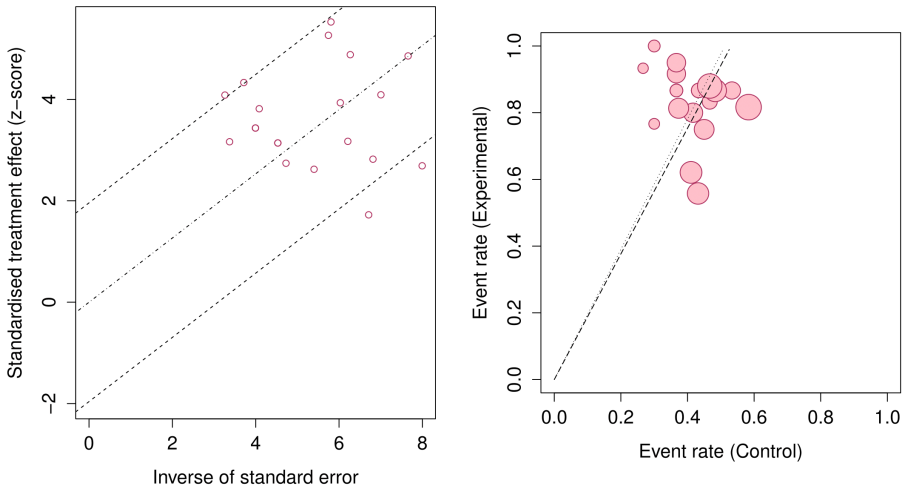

Figure 19: Left: Galbraith plot with added publication bias; Right: L'Abbe plot with added publication bias

4.4 Adding publication bias

4.4.1 Significant case

The next step was to add publication bias to the studies. The same study sizes from the previous section were used, except there was an increase in the probability of an event occurring in the treatment group. The smaller the study, the more publication bias which was added, as smaller studies are more prone to selective suppression than larger studies. From the trials with 30 patients, 0.3 was added, for trials with 60, 0.25 was added and for trials with 75, 0.2 was added. Nothing was added to the trials with 95 patients. Heterogeneity statistics were expected to increase significantly. The funnel plots are displayed in Figure 18.

There is clear asymmetry present, with the studies shifted towards the right hand side of the graph. The contour enhanced plot shows that the publication bias is present as the asymmetry is being caused by a lack of studies in a non-significant area of the funnel plot. Figure 19 displays a Galbraith plot and a L'Abbe plot.

The Galbraith plot has 3 or 4 studies outside the 95% confidence interval points (15-20% of studies) suggesting significant bias. This notion is reinforced by the L'abbe plot. The assumptions about baseline risk have been broken because all the studies lie between the x axis 0.2 and 0.6. This clustering is an indication of heterogeneity. The heterogeneity statistics came in at : I^2=45.6%,Q=34.92. The confidence interval for I^2 is [8.1% , 67.8%] with a p-value of 0.0143 showing that significant

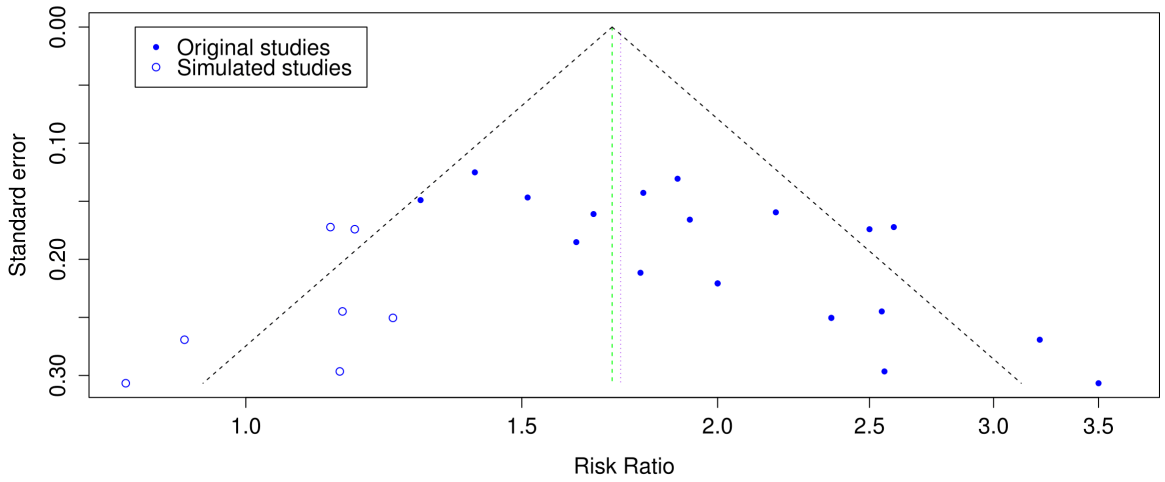

Figure 20: Trim and Fill with added publication bias

heterogeneity is present. A trim and fill adjusted funnel plot is displayed in Figure 20.

This is much more symmetrical. The random effects estimate before the trim and fill was 1.9526 with a confidence interval of [1.7485 , 2.1806]. Before the publication bias was added the random effects estimate was 1.5205 with a confidence interval of [1.3895 , 1.6638]. Clearly the publication bias is exaggerating the effect estimate. After performing a trim and fill the effect estimate decreased to 1.7344 with a 95% confidence interval [1.5387 ; 1.9549]. The trim and fill has lowered the random effects estimate to within 0.02 of its actual value. This method may have a low power but has made a huge improvement in this case. Note also that the confidence interval is smaller than the estimate where there was no publication bias, this is because 7 simulated studies have been added.

The overall heterogeneity statistic for I^2 after the trim and fill has increased from 45.6% to 61.8% but its confidence interval has decreased from [8.1% ; 67.8%] to [42% ; 74.9%]. The trim and fill technique has reduced asymmetry and has given a much better estimate of the 'true' random effects estimate. Although asymmetry has decreased, its confidence interval has been been slightly reduced. This example is an ideal example of when a trim and fill should be executed.

4.4.2 Significant case, less bias

The last section displayed an extreme version of publication bias which is very easy to spot using a funnel plot. However, in reality publication bias is rarely this extreme or obvious. In this section the bias will be reduced gradually to see at what point the publication bias is impossible to spot graphically using funnel plots. Previously 0.3, 0.25 and 0.2 were added to the smallest, second

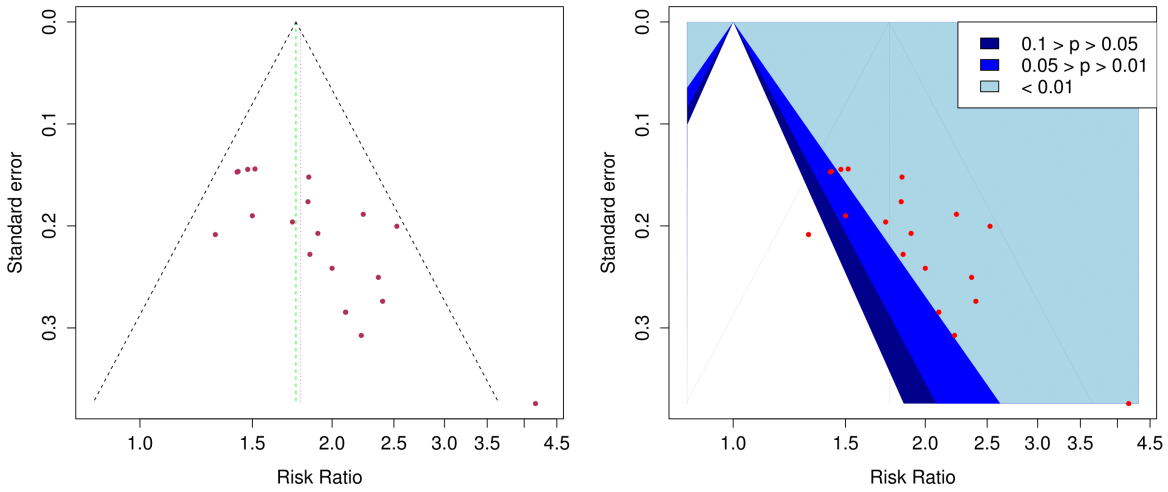

Figure 21: Funnel plots with publication bias halved

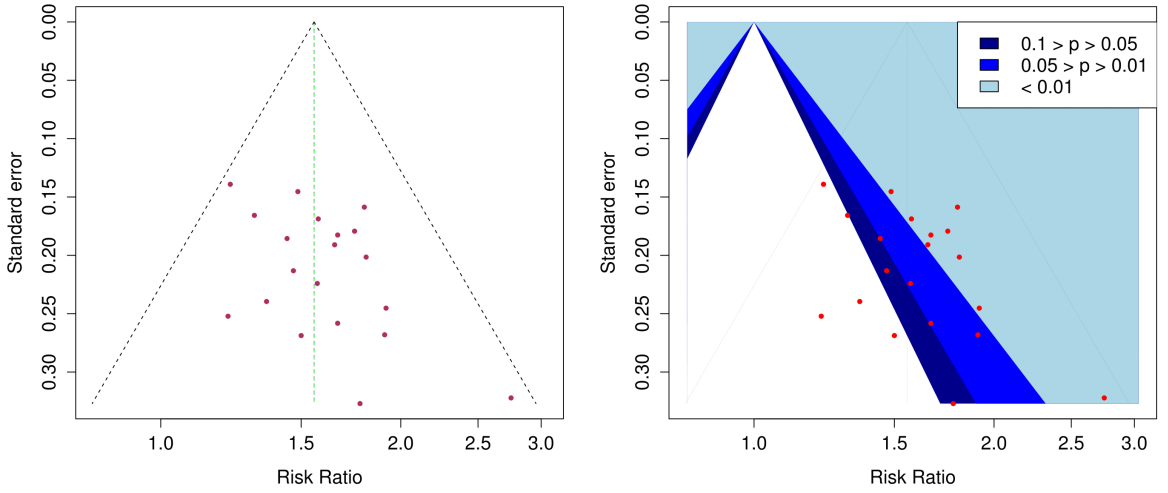

Figure 22: Funnel plots with publication bias halved again

smallest and third smallest trials. This was decreased by a half to 0.15, 0.125 and 0.1 to give the funnel plots in Figure 21.

There is still evidence of publication bias present, but its effect is less pronounced. The value of I^2 came in at 21.5% with a 95% confidence interval of [0% ; 54.3%]. Note that as the amount of publication bias being added is halved, the heterogeneity statistic is reduced by around a half too. The previous levels of publication bias added were reduced again by a factor of a half. Figure 22 displays the effect when 0.075, 0.0625 and 0.05 were added to the smallest, second smallest and third smallest trials.

It is becoming much more difficult to tell if there is publication bias or not. There still appears to be some hints but this could be down to chance. This suggests that the borderline values for detecting

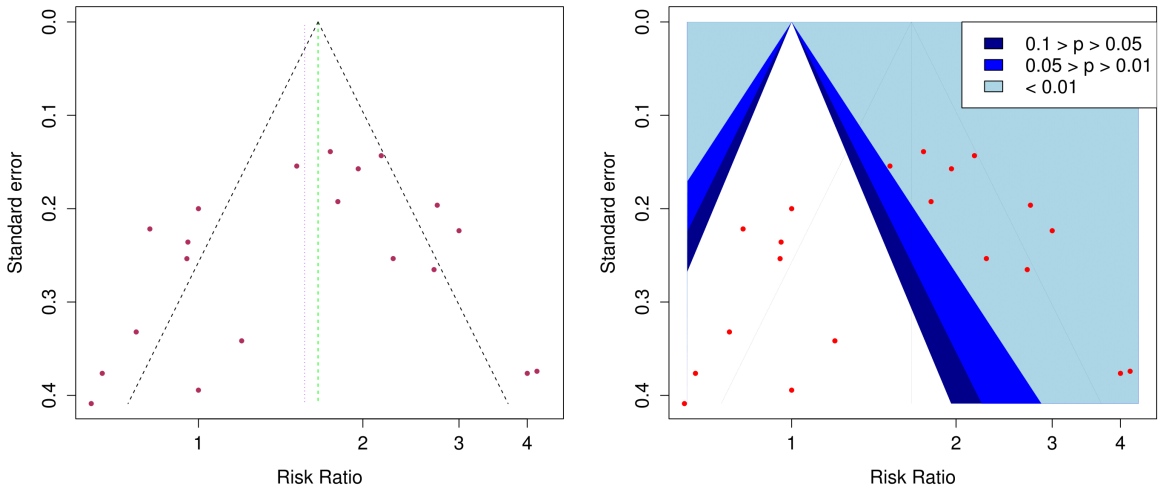

Figure 23: Funnel plots with insignificant publication bias

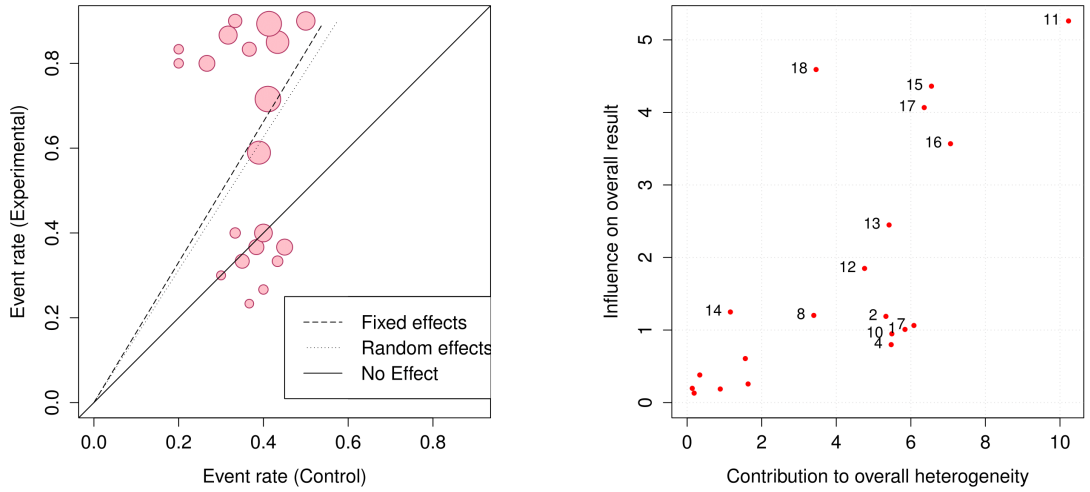

Figure 24: Left: L'Abbe plot with insignificant publication bias; Right: Baujat plot with insignificant publication bias

heterogeneity via funnel plots are around 0.075, 0.0625 and 0.05.

4.4.3 Adding publication bias (Insignificant case)

Publication bias is not always pushing the effect estimate one way or another, sometimes it is insignificant trials which are not published. This type of publication bias is the more likely of the two to occur (Deeks 2002). A problem arises in that it is difficult to spot using heterogeneity statistics alone, graphs must be used to assess what is causing the heterogeneity. To simulate this, half of the treatment group's event probabilities must be increased and half must be decreased, with this effect being more exaggerated in smaller studies. The funnel plots are displayed in Figure 23.

There appears to be some asymmetry caused by studies being shifted slightly to the left, but what's more striking is the lack of smaller sized studies present up the middle of the funnel plots. This, combined with so many studies in areas of high significance, suggests that insignificant studies have been suppressed. The random effects estimate here was 1.5650 and I^2 was 64.5% with a 95% confidence interval [44.4% ; 78.4%]. The L'abbe plot and Baujat plot are displayed in Figure 24.

The L'abbe plot shows clustering around the x axis, again indicating publication bias. The Baujat plot interestingly shows the splitting of the smaller studies with a large gap through its middle as the plots move diagonally upwards. This is another sign that it is insignificant studies causing the publication bias. Heterogeneity levels are much too high here and because it is insignificant studies causing the asymmetry, a trim and fill would not help. Results from this meta-analysis would be misleading, and so the meta-analysis should be carried out again but with unpublished studies included.

4.5 Identifying an outlier

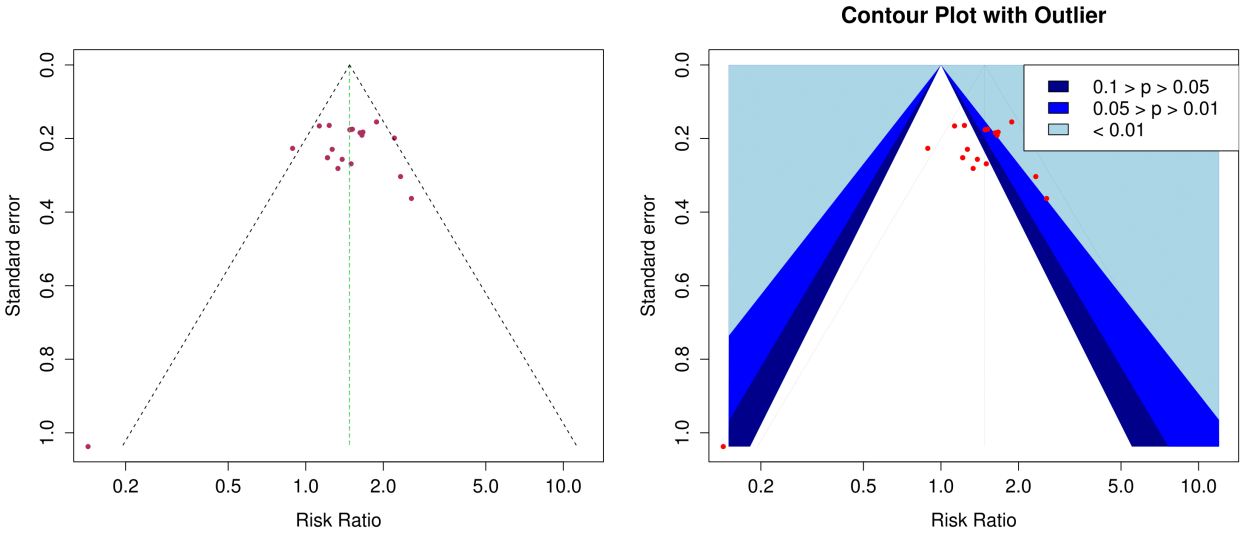

Figure 25: Funnel plots with outlier

Outliers can skew a meta-analysis and detecting them should be of high priority. The following meta-analysis will be done with an outlier present. Graphical techniques will be used to detect which study is the outlier so the researcher can remove it and re-estimate the effect size. The data set from 4.1.3 has an ordinary study with 30 patients changed into an outlying study. This outlying study has a large reduction in the probability of an event occurring in the treatment group. Figure 25 displays the funnel plots.

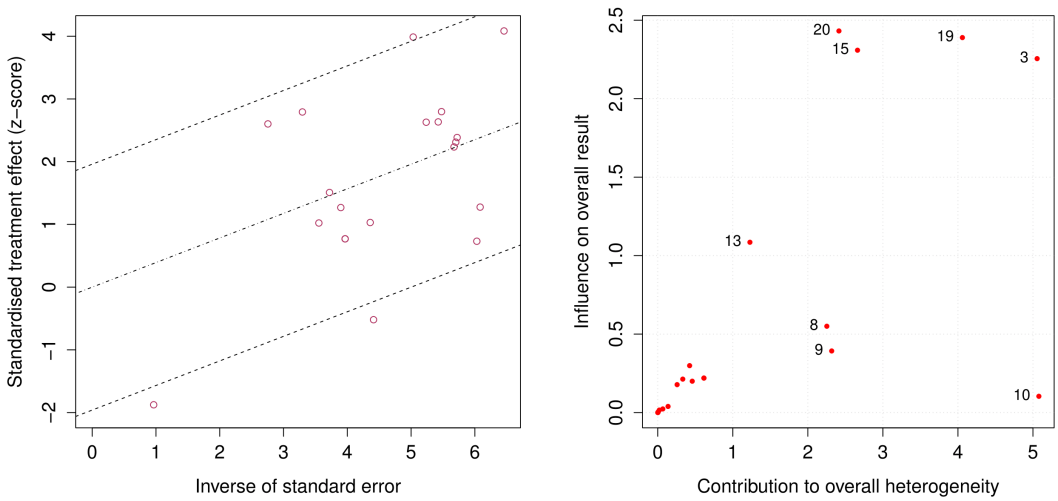

Figure 26: Left: Galbraith plot with outlier; Right: Baujat plot with outlier

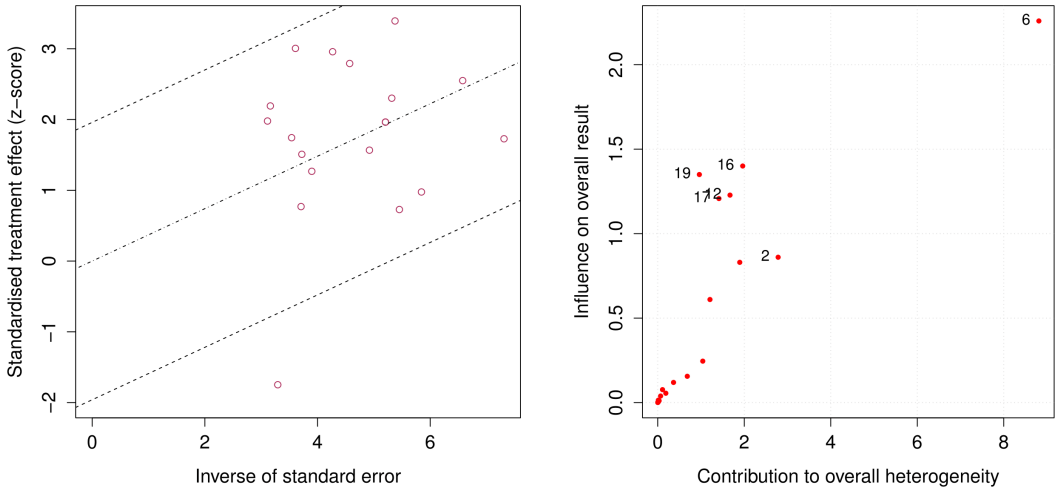

Figure 27: Left: Galbraith plot with outlier removed; Right: Baujat plot with outlier removed

The funnel plots show that there is a potential outlier on the bottom left in otherwise symmetric plots. Figure 26 displays a Galbraith plot and Baujat plot to investigate this further.

There are 3 studies outside the 95% significant area indicating heterogeneity, but apart from the potential outlier the funnel plots were otherwise symmetric and homogeneous. The Baujat plot shows that there is one point which disproportionately adds heterogeneity compared to its influence on overall result. This study, number 10, is the outlier and is removed. The random effects estimates and values for I^2 are recorded before and after removal. I^2 decreases from 49.1% to 0%. Effect estimates drop from 1.4741 to 1.3941. Figure 27 displays the Baujat and Galbraith plots after the outlier has been removed.

The Galbraith plot now has only 1 study outside the 95% confidence lines as expected (=5%) and

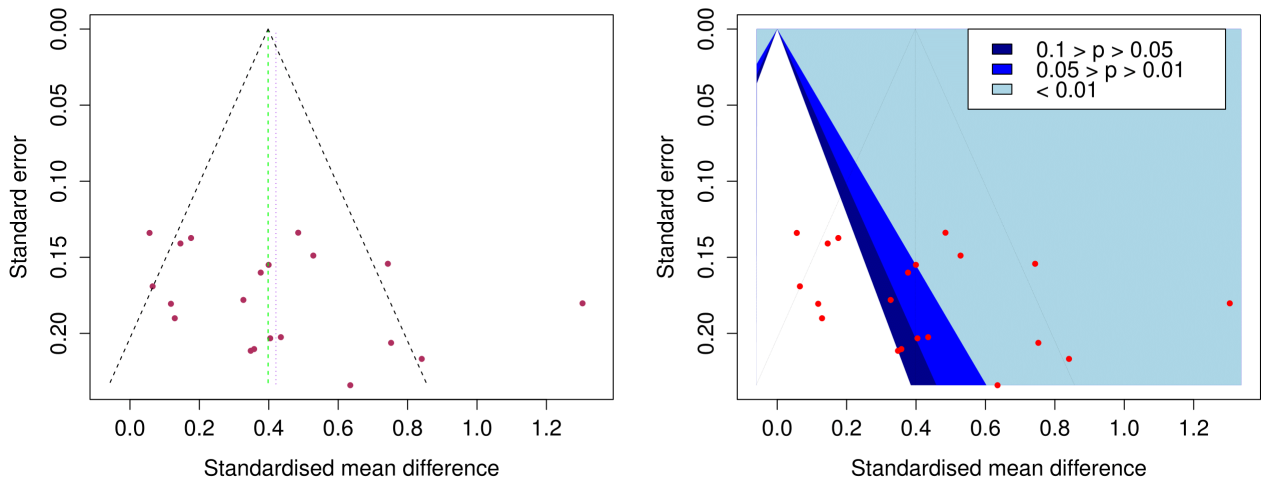

Figure 28: Funnel plots with continuous data

the Baujat plot has no studies with disproportionately high heterogeneity.

4.6 Extension to continuous data

The methods used for binomial data can easily be extended to continuous data as seen in section 3.8. As before, L'abbe plots only work for binomial data but all other graphical techniques work. In this next section the data simulated will not be the same format as the previous section. There are 20 trials and the trial sizes can range from 37 to 118 and the treatment group/control group ratios do not necessarily equal 1. This is because in reality patients are randomly allocated to either group. The treatment and control groups are then often slightly different sizes. In these data, treatment group and control group numbers will be within three of each other which will increase the heterogeneity in these meta-analyses.

4.6.1 Continuous data with no bias added

The first case being considered will be the case with no added publication bias. This funnel plot is expected to be symmetrical and any asymmetry which may be present is simply due to chance. Figure 28 shows the funnel plots for this data.

The plots are fairly symmetric with one fairly extreme value on the right hand side. There are 4 studies outside the 95% lines, but this could be due to the outlier. The value for I^2 here came out at 37.4% with a 95% confidence interval [25.5% ; 56.2%]. The outlier could be removed and effect

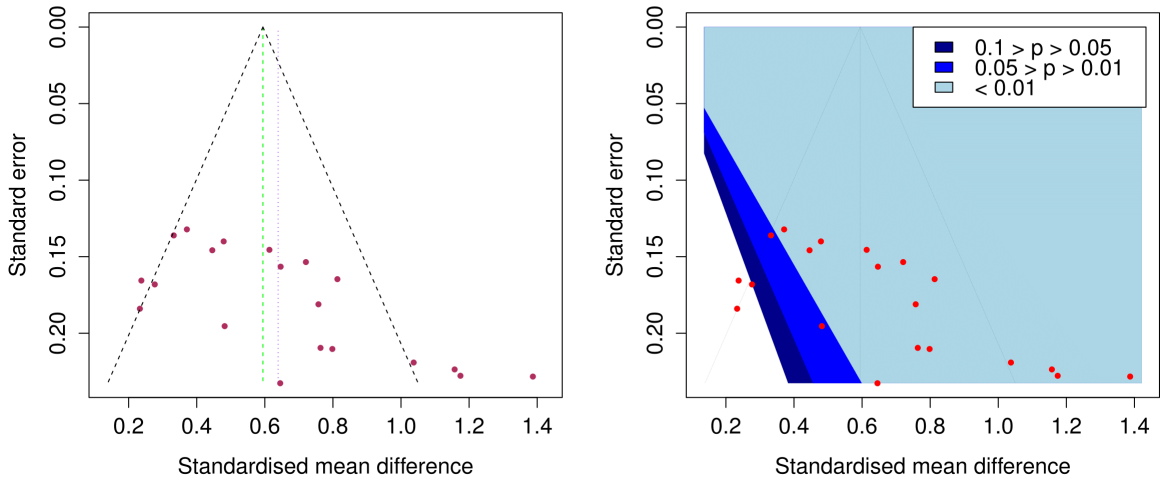

Figure 29: Continuous data Funnel plots with added publication bias

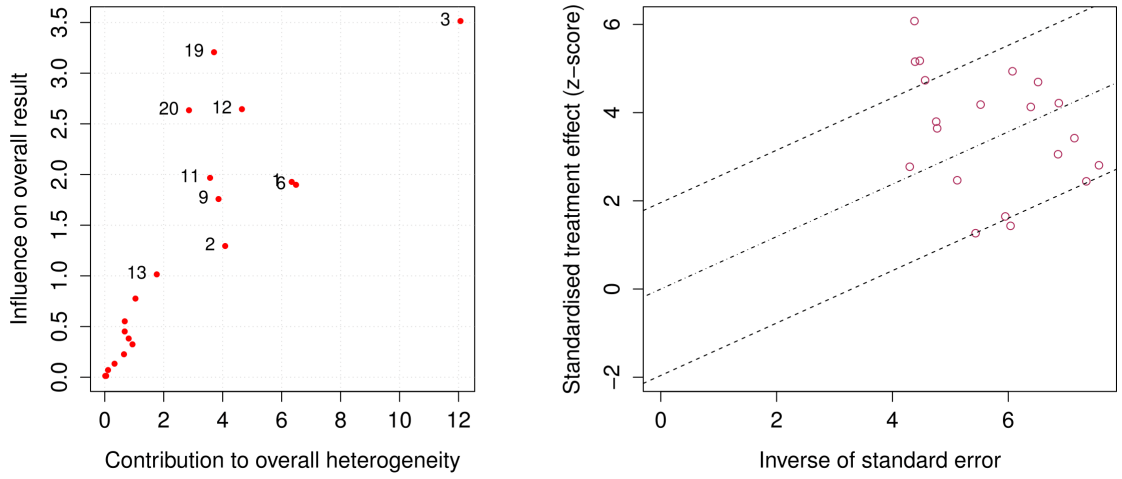

Figure 30: Left: Baujat plot for continuous data with added publication bias; Right: Galbraith plot for continuous data with added publication bias

re-estimated, but it doesn't seem to be causing that much of a problem. The random effects estimate came out at 0.4207. There is no further analysis needed here.

4.6.2 Continuous data with publication bias added

The publication bias being added is going to be that of negative findings not being published. To do this an equi-spaced sequence of length 20 ranging from 0.2 to 0 was added to the effect size in the treatment group. This shift allows for more publication bias in lower studies and less in larger studies as required. The funnel plots are displayed in Figure 29.

Smaller studies have shifted to the right causing asymmetry. This asymmetry is caused by 'missing'

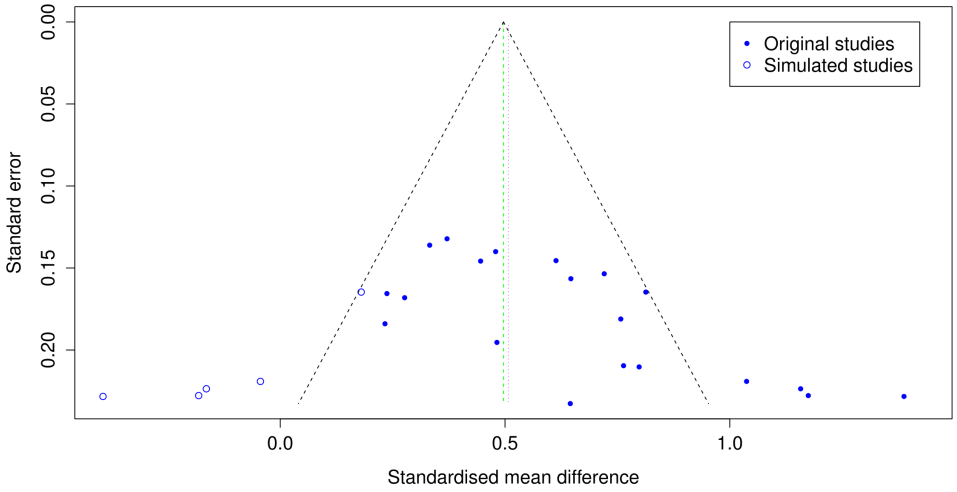

Figure 31: Trim and Fill for continuous data

studies in the bottom left corner in an area of non-significance. Clearly this is due to publication bias (section 3.7). Figure 30 shows the Galbraith and Baujat plot.

There does not appear to be any outliers in the Baujat plot as there are no points contributing highly to heterogeneity and contributing very little to influence on overall result. However the Galbraith plot shows 25% of the studies outside the 95% lines, indicating high levels of heterogeneity. Clearly publication bias is present and a trim and fill should be carried out, this is displayed in Figure 31.

The funnel plot now looks a lot more symmetric, however there is still a high proportion of studies outside the confidence lines. A total of 5 extra studies have been added. The random effects estimate has decreased from 0.6391 to 0.5076. This is closer to the value when no bias was added but is still a little high. The 95% confidence interval has increased slightly from [0.5090 ; 0.7691] to [0.3607 ; 0.6545]. This increase along with an increase in I^2 from 65.3% to 76.9% shows that although we have got a better treatment effect estimate, it has increased variability. Overall the benefits outweigh the drawbacks.

4.6.3 Identifying an outlier

Outliers can cause heterogeneity and misleading effect estimates. As before, graphs will be used to identify an outlier and then analysis can be repeated with the outlier removed. The outlier data set was made by subtracting 1 from one of the smaller data set's effect size. Already it appears that the study in the bottom left is a potential outlier. A Baujat plot and Galbraith plot are shown in Figure

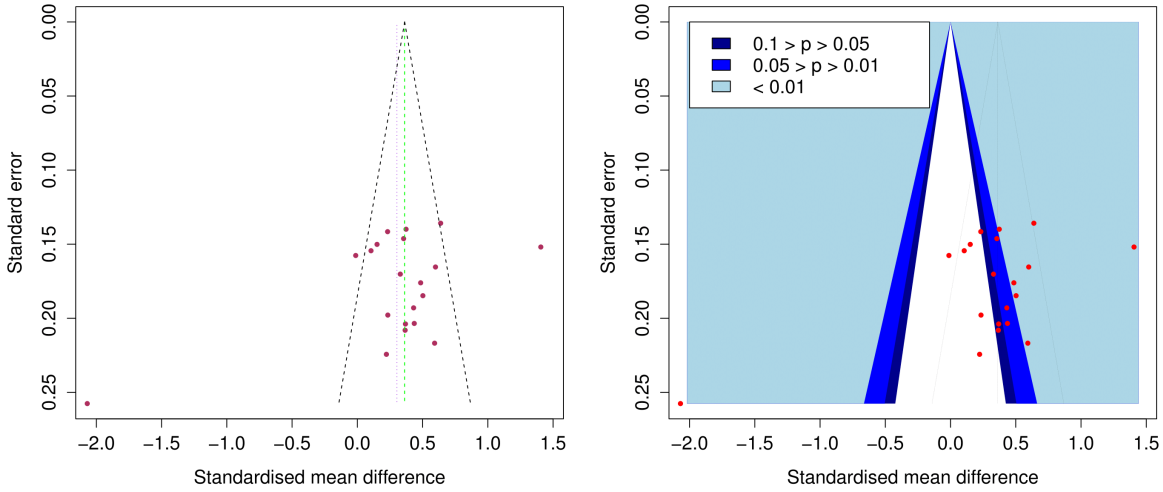

Figure 32: Funnel plots with an outlier for continuous data

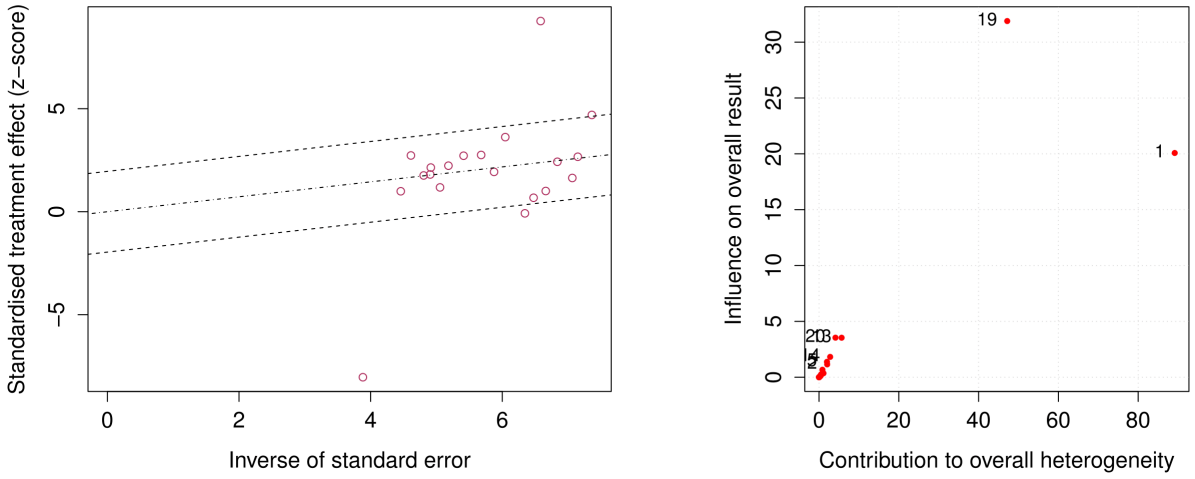

Figure 33: Left: Galbraith plot with an outlier for continuous data; Right: Baujat plot with an outlier for continuous data

33.

The Galbraith plot shows one study very far outside the 95% confidence lines which again could be a possible outlier. The Baujat plot shows one study (labeled '1' in Figure 33) which has higher level of heterogeneity than the other studies. This was the study which had 1 subtracted from its treatment effect and is the outlier. This should be removed and the treatment effect should be re-estimated. This is shown in figure 34 with a funnel plot and Baujat plot.

The outlier has been removed and the funnel plot is less asymmetric. The Baujat plot shows no indication of any outliers and the random effects effect estimate comes out at 0.3928, which is an increase from 0.3030.

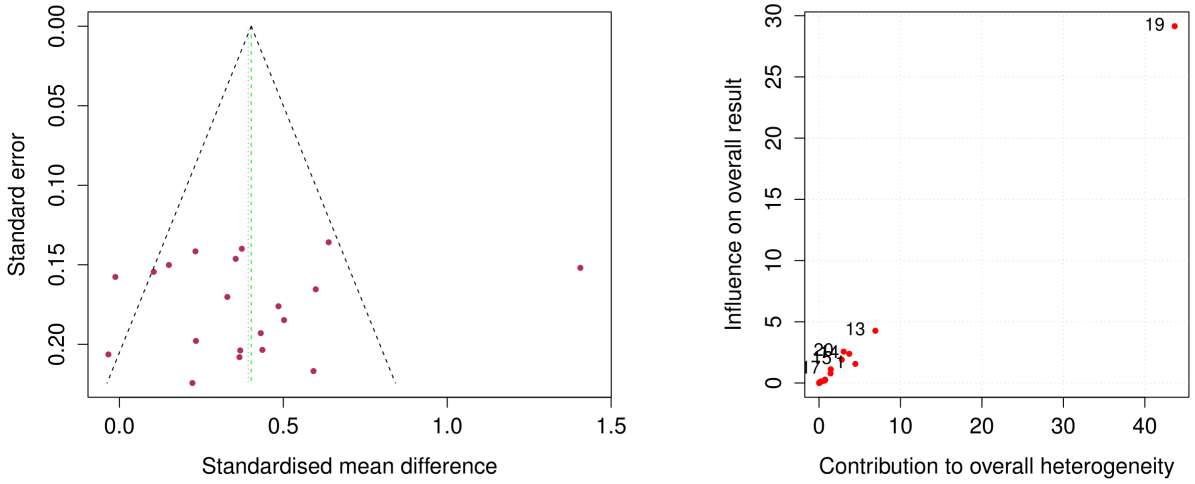

Figure 34: Left: Funnel plot with outlier removed; Right: Baujat plot with outlier removed

5 Bayesian meta-analysis

5.1 Bayesian meta-analysis

With an explosion in the use of meta-analyses, Bayesian meta-analyses methods are becoming increasingly popular within medical statistics. One of the main reasons for its recent popularity is advances in computers along with its more appealing nature. There are clear advantages with a Bayesian approach not present in frequentist statistics. The Bayesian approach is to combine a prior distribution with a likelihood, giving you a joint posterior function up to proportionality.

Some notation must be reconsidered for the multivariate case. μ must be defined as the common effect size in the fixed effect model and all other extensions are intuitive. This leaves us with the following model

$$T_i \sim N[\theta_i, \nu_i], \quad i = 1, ..., k \tag{19}$$

$$\theta_0 \sim N[\mu, T^2], \quad i = 1, ..., k \tag{20}$$

Bayesian methods differ from frequentist methods in that the data and model parameters are taken to be random quantities, and the likelihood function is thought of as defining the plausibility of the data given values of the model parameters (Sutton 2001). Prior beliefs are chosen based on external evidence from related studies. Usually the only two parameters of interest are the effect size and marginal posterior densities for these parameters, where marginal posteriors are found from integrating the joint posterior densities with respect to all of the other parameters. Difficulty can arise here with evaluation of high dimensional integrals, this can be made simpler with asymptotic approximation or conjugate models. Computational techniques such as Markov Chain Monte Carlo (MCMC) simulations are used more frequently due to modern day computational power. MCMC is a type of Gibbs sampler, and Gibbs samplers sample from posterior condition distributions to get marginal posterior distributions.

5.2 Dissemination of failed reviews

In some meta-analyses, conclusions may not be reached. Usually this is due to poor methodological design, but there are a variety of other reasons. Poor methodological design leads to study parameters which may be too varied to combine sensibly. Sometimes meta-analyses which consist of a small number of trials are also deemed inconclusive, even if a significant effect is found (Smith 1998). Abandoning the data found in these cases is a waste of collected information. A researcher should go back and recollect the information and conduct analysis again, but with a stricter methodological design.

5.3 Non-Normally distributed meta-analysis

Random effects and fixed effects meta-analysis assumes that effect sizes follow a normal distribution. However there is usually little justification for this assumption, effect sizes being normally distributed tend to be the exception rather than the rule. This is especially the case with meta-analyses that have a small number of studies. The impact of assuming a different distribution has been relatively under researched (Kontopantelis 2010).

There are a number of different effect sizes distributions which could be considered and a range of different parametric and nonparametric methods for effect estimation. To test this studies would be simulated with different distributions. These different methods would then be used to test if there is a significant difference in effect estimation.

5.4 Meta-regression

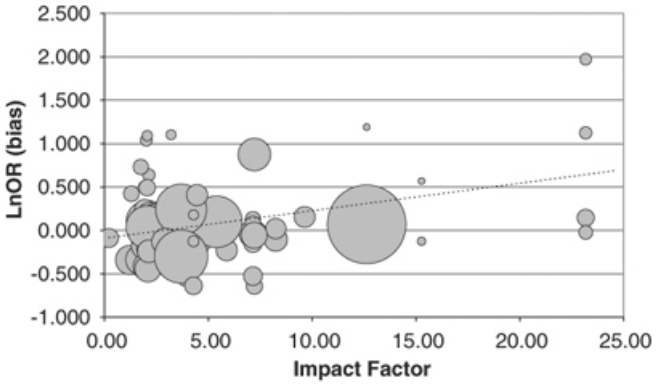

Figure 35: Meta-regression plot

Statistical heterogeneity is any variation which is not due to chance alone. This is mainly caused by different methodological design, something which is almost impossible to keep exactly the same. Meta-regression can be used to explain the characteristics involved in generating an effect size. It follows that if particular characteristics are causing heterogeneity, they can then be identified and progress can be made in reducing heterogeneity. Meta-regression requires visual representation and Figure 35 shows a meta-regression plot. A graphical presence is required to visualise how the regression coefficient is found in relation to study size and the distribution of information. It does however share a common problem in that the technique is not very effective when too few studies are considered (Thompson 2002).

The characteristics which explain the heterogeneity are known as *effect modifiers* and it would be reasonable to assume that all heterogeneity can be explained by them. The possibility of residual heterogeneity must also be entertained and for this reason only random effects meta-regression will be considered. The weighting of meta-regression uses the inverse of the variance, which is the same as with meta-analysis. Random effects and fixed effects meta-regression differs in that with fixed effects, variance is the within study variance. For random effects meta-regression it is the sum of between study variance and the within study variance. The test for overall heterogeneity has a low power, so if a meta-analysis has a nonsignificant heterogeneity statistic, there may be characteristics contributing to it. A meta-regression should still be performed in the absence of a significant result to check for effect modifiers.

Meta-regression does still have its limitations. The relationships observed are across all the trials and are not allocated to randomisation. The effect modifiers are then likely to be affected by confounding. Relationships found between trials may not be found between patients in a single trial. This phenomenon is known as ecological confounding. Investigating ecological confounding requires individual patient data. Even with its drawbacks, meta-regression is a tool which is very important for finding and eradicating bias. It follows naturally from this study in detecting bias.

6 Conclusion

The idea behind combining results extends back as far as 1904, but for a long time it was an underused technique. In 1976 Glass coined the term meta-analysis and sparked interest in the subject area through its applications in social sciences. From then on it became popularised for mathematicians to work on methods to perfect meta-analysis into the useful tool it is today.

A meta-analysis can be conducted no matter how many studies are involved. However, an issue arises when too few studies are combined. This means that in order to get a significant result there must be a large overall treatment effect. When more studies are included, it becomes easier for a researcher to say that the effect size is due to the treatment, rather than being simply down to chance. This is the idea behind statistical power, and improving the statistical power to detect an effect is one of the main advantages of meta-analysis.

There are two models considered in this study, fixed effects and random effects. Fixed effects models assume one true effect size and any variation from this value is simply due to chance. The only source of variance is within study variance, meaning any inference gathered with this model can only be said to hold for that particular set of studies. The absence of between study variance is a situation which happens only rarely, if ever. However, this model can still be useful when dealing with a group of studies with low between study variance. Random effects models assume that the effect sizes are a random sample from the relative distribution of effect sizes. There are two sources of variance here, within study variance and between study variance. Inference gathered with this model can then be said to hold globally. This is a more realistic situation and so random effects models should be used in almost all situations.

Researchers usually want to get a significant treatment effect in both meta-analyses and clinical trials. This introduces sources of bias. The problem with bias is that it contributes to the heterogeneity of a meta-analysis and may give misleading effect estimates. There are many types of bias, but the most problematic is publication bias. Detecting bias and pinpointing which type of bias is present is a necessity of meta-analysis. Graphical techniques are a good way for not only spotting bias, but also detecting which type of bias is present. To numerically measure the amount of heterogeneity present there are a range of statistics available. The best of which is I^2 because it describes the percentage of variance that is not down to chance and adds a measure of consistency to the results. I^2 does not change over the degrees of freedom and it measures heterogeneity as a percentage.

Graphical techniques are of great importance in meta-analysis. Some of them are good at spotting particular problems, but the one which contained the most information is the funnel plot. Funnel plots show bias and outliers and with the aid of contours can even help decipher what kind of bias is present. Each graphical technique used in this study has a purpose and so they are all important.

Simulating studies was perhaps the most informative part in relation to both measuring heterogeneity and testing the power of these graphical techniques. Varying the number of trials had relatively little effect when the number of patients was the kept the same. Varying the size of the studies on the other hand also had a small effect. However, any heterogeneity observed would not be significantly changed by varying either of these factors. Simulating a data set with publication bias significantly increased heterogeneity. Even the slightest changes increased the heterogeneity statistic and made the funnel plots asymmetric. Funnel plots were very good at picking up publication bias, but there was a threshold where it was difficult to infer if it was publication bias or chance that was causing the asymmetry observed. If a researcher was unsure with the levels of heterogeneity after a funnel plot, a L'abbe plot or Galbraith plot can be very useful in cementing or disproving this notion. If heterogeneity is believed to be a present a trim and fill can help adjust the effect estimate. This method has a low power and works on the assumption that a funnel plot will be symmetric in the absence of bias, but it does improve the estimate in these conditions. Deciding whether to use a trim and fill depends on the levels of bias. If there is a lot of bias it makes sense to do one, if there is relatively little bias it's better to proceed without.

There is another type of publication bias which is actually more common, where insignificant studies are not published but significant studies in either direction still are. Most of the plots in this study are fairly poor at picking this up except for the funnel plot. The funnel plot will have a gap in the middle. Contour enhanced funnel plots play a role here because they can show whether it is bias or poor methodological design causing asymmetry present in a funnel plot.

Identifying an outlier can be tricky, especially with funnel plots because the scale can make it difficult to gauge whether a point is really far out or not. If a funnel plot shows signs of containing an outlier the best plot to use is a Baujat plot. They are more likely to contribute more highly to heterogeneity than influence on the result and so should appear low and to the right. All obvious outliers should be highlighted and removed before re-estimation.

The extension to continuous data is relatively straightforward. All of the graphical methods used for binary data also work for continuous data with the exception of the L'abbe plot. All inferences can be taken in the same way but different summary measures must be used. For binary data the best summary measure was risk ratio because it was relative, easy to understand and had good mathematical properties. For continuous data the best summary measure was the standardised mean difference as it was the only relative term.

Although it has made lots of advances in recent years, meta-analysis still has a few problems. The first being the presence of avoidable heterogeneity. There will always be some unavoidable heterogeneity caused by differences in trial design and assessment, but it should end there. The presence of publication bias can be stopped by publishing all findings, regardless of their result. Institutions should persuade all research to be published, and those who do not should face repercussions. There are also some poor meta-analyses being performed from bad clinical trials, and for this reason many researchers will side with a single large clinical trial over a meta-analysis of several smaller trials. The graphical tools displayed in this study were only a few of the most important techniques for detecting heterogeneity, and there are many others which have not been discussed. These tools and an increased standard of analysis has led to meta-analysis becoming one of the most implemented techniques in genetics, drug testing and anything involving a clinical trial.

References

Anzures-Cabrera, J. and J. Higgins

2010. Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, 1(1):66–80.

Borenstein, M., L. V. Hedges, J. Higgins, and H. R. Rothstein

2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111.

Cohen, J., P. Cohen, S. G. West, and L. S. Aiken

2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Copas, J. and J. Q. Shi

2000. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1(3):247–262.

Deeks, J. J.

2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in medicine*, 21(11):1575–1600.

Egger, M. and G. D. Smith

1998. Meta-analysis bias in location and selection of studies. *Bmj*, 316(7124):61–66.

Egger, M., G. D. Smith, M. Schneider, and C. Minder

1997. Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109):629–634.

Else-Quest, N. M., A. Higgins, C. Allison, and L. C. Morton

2012. Gender differences in self-conscious emotional experience: a meta-analysis. *Psychological bulletin*, 138(5):947.

Glass, G. V.

1976. Primary, secondary, and meta-analysis of research. *Educational researcher*, Pp. 3–8.

Hartung, J. and G. Knapp

2001. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24):3875–3889.

Hedges, L. V. and I. Olkin

1980. Vote-counting methods in research synthesis. *Psychological bulletin*, 88(2):359.

Higgins, J. and S. G. Thompson

2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558.

Hunter, J. E. and F. L. Schmidt

2000. Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, 8(4):275–292.

Kontopantelis, E. and D. Reeves

2010. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical methods in medical research*, P. 0962280210392008.

Montori, V. M., M. Smieja, and G. H. Guyatt

2000. Publication bias: a brief review for clinicians. In *Mayo Clinic Proceedings*, volume 75, Pp. 1284–1288. Elsevier.

Munafò, M. R., T. G. Clark, and J. Flint

2004. Assessing publication bias in genetic association studies: evidence from a recent meta-analysis. *Psychiatry research*, 129(1):39–44.

PAI, M., M. McCULLOCH, J. D. GORMAN, N. PAI, W. ENANORIA, G. KENNEDY, P. THARYAN, and J. M. COLFORD Jr

2004. Clinical research methods. *National Medical Journal Of India*, 17(2).

Peters, J. L., A. J. Sutton, D. R. Jones, K. R. Abrams, and L. Rushton

2008. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of clinical epidemiology*, 61(10):991–996.

Peters, J. L., A. J. Sutton, D. R. Jones, K. R. Abrams, L. Rushton, and S. G. Moreno

2010. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):575–591.

Rosenthal, R.

1979. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638.

Simpson, R. and K. Pearson

1904. Report on certain enteric fever inoculation statistics. *The British Medical Journal*, Pp. 1243–1246.

Smalley, W., D. Shatin, D. K. Wysowski, J. Gurwitz, S. E. Andrade, M. Goodman, K. A. Chan, R. Platt, S. D. Schech, and W. A. Ray

2000. Contraindicated use of cisapride: impact of food and drug administration regulatory action. *Jama*, 284(23):3036–3039.

Song, F.

1999. Exploring heterogeneity in meta-analysis: is the l'abbe plot useful? *Journal of clinical epidemiology*, 52(8):725–730.

Sterne, J. A., D. Gavaghan, and M. Egger

2000. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology*, 53(11):1119–1129.

Sterne, J. A., A. J. Sutton, J. Ioannidis, N. Terrin, D. R. Jones, J. Lau, J. Carpenter, G. Rücker, R. M. Harbord, C. H. Schmid, et al.

2011. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Bmj*, 343.

Su, R., J. Rounds, and P. I. Armstrong

2009. Men and things, women and people: a meta-analysis of sex differences in interests. *Psychological bulletin*, 135(6):859.

Sutton, A. J. and K. R. Abrams

2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4):277–303.

Thompson, S. G. and J. Higgins

2002. How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine*, 21(11):1559–1573.