https://doi.org/10.26219/heal.aueb.6071

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

"Introducing meta-analysis methods for formulating, evaluating and combining information from medical studies"

By

Nikoletta G. Antoniou

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

July 2014

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

«Εισαγωγή στις μεθόδους της μετα-ανάλυσης για τη διαμόρφωση, την αξιολόγηση και τον συνδυασμό πληροφοριών από ιατρικές μελέτες»

Νικολέττα Γ. Αντωνίου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα

Ιούλιος 2014

https://doi.org/10.26219/heal.aueb.6071

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Professor V. Vasdekis, for his assistance in the completion of this dissertation and for his patience.

Furthermore, I am grateful to the Faculty and staff of the program of Master of Science in Statistics of the Athens University of Economics and Business.

I would also like to thank my family and friends for all their support and encouragement.

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

ABSTRACT

Nikoletta Antoniou

"Introducing meta-analysis methods for formulating, evaluating and combining information from medical studies"

July 2014

Since 1980, an upsurge in the application of meta-analysis in medical research has been observed. Along with the usage of meta-analysis, improvement and developments in statistical methodology have also been accomplished. These developments have to do with the need of using evidence to justify the medical results along with the need of having reliable summaries. Most of the meta-analyses within the clinical research have been conducted on randomized controlled trials.

The form and the amount of the available data used for a meta-analysis may vary. We may have individual patient data or summary from p-values in order to conduct our meta-analysis. That is the reason why many methodologies were developed in order to run a meta-analysis, along with their practical implementation. In this thesis, our aim is to present as many of these methodologies as possible. Furthermore, we would like to place the general framework of the statistical methods that are used in each approach.

It is a common fact that most of the meta-analyses are retrospective and based on published summary statistics from reports of individual clinical trials. However, we should not overlook the results of a prospective meta-analysis neither the advantages of using an individual patient data. In this thesis, techniques from both prospective and retrospective analysis are included, as well as methods based on individual patient data and summary statistics.

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

ΠΕΡΙΛΗΨΗ

Νικολέττα Αντωνίου

«Εισαγωγή στις μεθόδους της μετα-ανάλυσης για τη διαμόρφωση, την αξιολόγηση και τον συνδυασμό πληροφοριών από ιατρικές μελέτες»

Ιούλιος 2014

Από το 1980, έχει παρατηρηθεί μεγάλη χρήση της μεθόδου της μετα-ανάλυσης στον κλάδο των κλινικών ερευνών. Όμως, μαζί με τη χρήση της μετα- ανάλυσης, παρατηρήθηκε και μεγάλη ανάπτυξη καθώς και βελτίωση των μέχρι τότε στατιστικών μεθοδολογιών. Όλη αυτή η ανάπτυξη έγινε με σκοπό να βασίζονται τα αποτελέσματα των ερευνών σε τεκμηριωμένα μαθηματικά στοιχεία αλλά και σε αξιόπιστα αποτελέσματα από μεριάς στατιστικής ανάλυσης.

Τόσο η μορφή όσο και η ποσότητα των δεδομένων που χρησιμοποιούνται σε κάθε μετα- ανάλυση είναι πιθανό να διαφέρουν. Είναι δυνατόν να χρησιμοποιηθούν απευθείας δεδομένα από τους συμμετέχοντες στην έρευνα, όμως μπορεί και να χρησιμοποιηθούν πίνακες με αποτελέσματα από προηγούμενες μελέτες. Αυτός είναι άλλωστε και ένας από τους λόγους που αναπτύχθηκαν διαφορετικές μέθοδοι στον τομέα της μετα-ανάλυσης. Επιπρόσθετα, μια μετα- ανάλυση μπορεί να διακριθεί ανάλογα με το αν έχει αναδρομική μορφή ή όχι, αν και είναι γεγονός πως οι περισσότερες μετα-αναλύσεις οι οποίες διεξάγονται παγκοσμίως έχουν αναδρομική μορφή.

Σκοπός αυτής της διπλωματικής εργασίας, είναι η παρουσίαση όσο το δυνατόν περισσότερων μεθοδολογιών της μετά-ανάλυσης, με παράλληλη παρουσίαση του πλαισίου των στατιστικών μεθόδων που χρησιμοποιούνται σε κάθε μια από αυτές.

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

TABLE OF CONTENTS

https://doi.org/10.26219/heal.aueb.6071

LIST OF TABLES

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

CHAPTER 1

INTRODUCTION

1.1: Introduction to meta-analysis

In research we really do not trust any outcome or any results until it is being replicated several times. When we have several experiments showing us similar outcomes, we need to combine the evidence. Meta-analysis does that!

Meta-analysis has to do with the methods that focus on combining the outcomes from different studies hoping to identify a certain pattern among them in order to highlight a result. Meta-analysis can be defined as the quantitative review and synthesis of the outcomes of related plus independent studies. In 1976, Glass [1] gave the definition of Meta-analysis as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings". At first, the term of Meta-analysis was used in social science but soon it has been adopted by other sciences too, proving to be extremely popular.

The goals of meta- analysis can be several. First of all, the possibility of combining information and results from different studies gives the opportunity of having more, statistical powerful, results which are able to lead in the detection of a treatment effect in comparison with the results of one and only study. Moreover, there is a possibility that several studies can lead to conflicted outcomes. In that case, a meta- analysis can be used to estimate an overall effect.

In the world of medicine, meta- analysis was widely spread in the 1980's. However, some of the techniques have been used long before the 80's. In fact, Pearson in 1904 used applications of meta- analysis in order to summarize the correlation coefficients from studies of typhoid vaccination. Furthermore,

Cochran and Yates in 1938 used the combination of estimates from different agricultural studies. Tippet and Fisher, in 1931 and 1932 respectively, presented different methods for combining p values.

In medical studies, there was a great increase in the number of meta-analyses in 90's. That increase was primarily due to the importance of evidence-based medicine[1] along with the need of reliable and strong based summaries of the clinical research. The combination of the systematic review of all the evidence from the relevant studies and meta- analysis provides a quantitative summary of the results. Sometimes, the systematic review includes meta-analysis. However, that is not always possible either because of lack of evidence, or because of the appearance of unexpected inconsistencies among the relevant studies.

In pharmaceutical studies, the scientists can use meta- analysis to summarize the results of two or more trials according the efficiency of a drug application or to identify some less frequent outcomes according the safety evaluation of each drug. However, meta- analysis' results cannot replace the efficiency from the results of each trial. There is a possibility of different statistically and clinically significant results. In that case, the scientists must take into account the clinical trials in order to lead in different results.

In 1999, Fisher examined the two conditions regarding the substitution of two controlled trials by a large one. The first condition relates with the strength of evidence for efficiency of each trial. Fisher proved that we have the same power of results if we have two controlled trials with statistical significance at the two sided approximately 5% for each trial, and one large trial with statistical significance at the two sided 0.125% level. The second condition has to do with the evidence of replicability. Every meta-analysis includes two or more trials, so the analysts must be in position to evaluate the consistency of the individual

[1] Evidence-based medicine is defined as "integrating individual clinical expertise with the best available external clinical evidence from systematic research" [2]

results. If there are inconsistencies, then we must examine the extent of this phenomenon in order to choose the right model for the meta-analysis.

In order to say that meta-analysis can provide useful outcomes, it must answers to the following points according to the guide published by the Committee for Proprietary Medicinal Products in 2001:

- Meta- analysis must give an overall estimation of the treatment effects of the trials.
- Those results must appear in pre-specified subgroups of patients.
- Must give a more powerful outcome than the individual trials.
- Must evaluate the safety in every subgroup of patients or possible rare side effects of the treatment.
- To estimate the relationship between the dose and the response.
- To compare and to evaluate conflicted results among several studies.

Chalmers and Lau in 1993 pointed out the importance of meta- analysis before any clinical trial [3]. It is important for the experts to know the number of patients that must be recruited for the study. Also, it is important to know if the critical questions are answered by the existing studies or there is a necessity for a new one. Meta- analysis answers all those questions. They note that meta-analysis gives the possibility to the experts to specify what data they need for their research.

In order to conduct a meta- analysis a team of scientists is needed. That team must include both statisticians and medical experts. Statisticians provide all the technical knowledge in order to perform the meta- analysis, and the medical experts are needed to identify the trials, to define the criteria for the trials, to define the existence of heterogeneity on the results as long as interpreting the outcomes.

As we have already pointed out, meta- analysis is mostly used in the field of medical research. However, over the past decades there has been a breakthrough in the development of statistical methods used for conducting meta-analyses.

1.2: Retrospective and Prospective meta-analyses

In order to perform a meta- analysis, a literature review must to be undertaken and then be summarized. The sources to be searched include published and unpublished literature, uncompleted research reports and work in progress. The meta-analyst first begins with searches of bibliographic reports and abstract databases, which provide information for published reports.

Often, meta-analyses are performed retrospectively on studies which have not been planned with this on mind. Furthermore, many have been based on summary statistics which are extracted by published papers. However, there are a number of problems arising by that which can affect the validity of the meta-analyses.

A huge limitation is that meta- analysis can include only studies for which the data is retrievable. If meta- analysis includes only published studies, then there is the possibility of published bias, whereby the bias arising from the probability of a study to be published depending on the statistical significance of the results. Moreover, even when a study is published, the analyst is able to choose only the statistically significant results among the overall ones. However, if the outcomes of interest have not been recorded or defined in the same way in each trial then there is a problem to combine them. Even when identical outcomes have been recorded in each trial but the way they have been calculated and summarized is different then there is still a problem of combining them, especially regarding to the choice of the subjects included or the mechanism used for the missing values. All the above disadvantages can be overtaken if the analysts obtain the data from all the studies included, but that requires time and effort.

Basically, the goal of a meta- analysis is to estimate the differences between two treatments and to conclude, if that is possible, which is the best. To do that, firstly the analyst must choose an appropriate measure of the treatment difference, for example the difference in means if we have normally distributed

data or log- odds for a binary one. In a retrospective meta-analysis the studies may vary in many things, such as the design, the population, the treatment regimen, the quality and the outcome. Considering those differences, it is a logical thought to say that the true treatment difference will not be the same in all trials. In other words, there will be heterogeneity between trials. According to Thompson [4], the heterogeneity must be taken in serious account on the overall outcomes. Also, there must be great care when the analyst selects the trials he will include in the meta-analysis and to the interpretation of the results.

Prospectively planning a series of studies with purpose to combine the results in a meta- analysis has clear advantages, since many of the problems occurred by retrospective meta- analysis disappear. The individual trial protocols can be designed to be identical with regard to the collection of data to be included in the meta- analysis, and individual patient data can be made available [5].

1.3: Fixed Effects versus Random Effects

The first step to conduct a meta- analysis, as a primary research, begins with a well- formulated question and design. What do you want to do? Do you want to validate results in an overall population or you want to lead to new studies? What are the operational definitions or the target population? What types of designs will be included in the research?

All the above questions can be answered through the methods of review, the models of statistical inference the analyst will use and the interpretation of the outcomes. The topic usually has been discussed in the context of a meta-analysis, where the data consist of trial estimates of the treatment difference along with their standard errors. If we are interested on making conclusions for the very populations we have sampled, then we consider to have fixed treatment levels. The only concern or uncertainty, we may have, arises from the results from sampling of people into studies. This type of variation can be considered to

be a within-study variation, which is a function of the number of patients in the primary study and the variability in the patient responses within the primary studies. In that case we have a fixed- effects model.

In the fixed-effects model, we consider to have the same true treatment for all trials. The standard error of each trial estimate is based on sampling variation within the trial. Generally speaking, we can say that the inferences are similar to those made when we conduct an analysis of variance (ANOVA), when we do not have inter- study variation in the mean outcome. The basic idea in fixed-effects model is that other levels of treatments are alike to those in the sample of the primary studies resulting that inferences will be the same. The standard error of each trial estimate is based on sampling variation within the trial.

On the other hand, in case when the population has different characteristics or when different effects are allowed, then a random-effects model would be more appropriate. The intuition underpinning random- effects models is that because there are many different approaches to conducting a study by perturbing the design in a small way, then there are many different potential treatment effects that could arise [6]. This model has the same concept with the ANOVA model in which there is inter-study variation in the mean outcome. Furthermore, in that model the population is the one in which there are infinitely many possible populations. The true difference in each trial is assumed to be a realization of a random variable, which is usually considered to be normally distributed. The consequence of that is that the standard error of each trial estimate is increased due to the addition of this between-trial variation [5].

Both models have advantages and disadvantages and it is rather difficult most of the times to know which model is more appropriate for the meta-analysis someone wants to conduct. Firstly, as we have already mentioned, extremely important is the universe we will use for our study. In some cases the universe may be too small to conduct a fixed-effects model or too big for a random-effects model resulting a study without a practical importance.

https://doi.org/10.26219/heal.aueb.6071

Furthermore, usually the estimation of the treatment difference along with the confidence interval based on a fixed-effects model provide useful summary of the results. The problem is that those results characterize only particular trials included in the meta-analysis without giving the total information for determining the difference in effect that can be expected in general for patients.

On the contrary, the random- effects model allows that between- trial variability giving the possibility for an overall estimation along with the account of its standard error. Naturally, anyone can say that this model has the advantage of producing more general results, so it can be consider being a more appropriate model for a meta-analysis. Unsurprisingly, there are some concerns regarding the use of the random-effects model in practice. First of all, this model has the assumption that the results from the trials which are included in the meta-analysis are representative of the results obtained from the total population of the treatment centers. In reality however the selection of the centers which participate in clinical trials is not at random. Moreover, when there are few trials included in the meta- analysis, it may wrong to try to fit a random-effects model as the calculation of the between-study will be unreliable due to that small number of trials participating. In fact, when there is only one available trial then we can only use the fixed-effects model for our analysis.

As we have already mentioned, the heterogeneity plays an important role in meta-analysis. If there is no heterogeneity between trials, which is rather rare, then both models will provide the same results. But since that is difficult, we must point out that the heterogeneity increases the standard error of the total estimate from the random effects model comparing to the fixed-effects model. The difference between the total estimations from the two models depends to a great degree on the magnitude of the estimates from the large informative trials in relations to the others [5]. In other words, if we have a meta-analysis based on one large study with small positive estimates along with more smaller studies with large positive estimates, then the total estimate derives from the random-effects model will be larger than the one from the fixed-effects model. In case we

want to use a random-effects model no matter what, we must be more conservative, meaning we must use a larger number of patients to demonstrate a significant treatment difference comparing to the fixed-effects model.

It could be useful to perform every time both models and to compare the results. The existence of great differences at the end indicates the need of further investigation. Those differences may derive from the variability in study quality, the difference in the study population or from the difference in study protocols.

1.4: Individual Data and Summary Statistics

There is a great diversity in the form or the amount of data we can use in a meta- analysis. There is the data which is available for all patients individually with a common measure used in all studies. But, on the other hand, there is data for which the only thing we know is the p- value associated with the test of a treatment difference, or even worse case, a statement in a published paper that p-value is smaller or not than 0.05. However, we may have summary statistics from published papers, individual patient data founded on similar but not identically defined outcome measures, or a mixture from both summary statistics and individually patient data.

A meta-analysis based on individual patient data can provide more reliable and comprehensive results than one based only on summary statistics from published papers. Such an analysis can be enriched with a standardized approach being taken from the extraction of the data and from handling of the missing data. Furthermore, the existence of the data, especially when the data provides information such as age, gender or disease severity, can gives as the opportunity to explore the relationship between these and the treatment difference. For an ever more successful analysis, we must devote time for planning, data collection and analysis stages.

Pharmaceutical statisticians are often in a good position to perform a meta-analysis on individual patient data since, almost every time they have access to the original data. Even in a retrospective data- analysis, the data from various trials is electronically stored in databases. Other than pharmaceutical industry, may face problems on having access to the original data. More details for practical issues concerning such a task can be found in Steward and Clarke (1995) in their paper resulting from a workshop taken place by Cochrane's working group [7].

Meta- analysis based on individual patient data have, as it is expected, clear advantages. But there are disadvantages too. All this processes is not only time consuming but also costly. Plus, a situation may come on surface in which additional resources needed to obtain individual patient data are not available or cannot be justified. Even when the plan is to select the data we want, it may not be possible to be obtained from all relevant studies. For those reason many times we conduct meta-analysis using summary statistics.

In case that meta-analysis wants to provide an overall estimate of treatment difference, an individual trial can be included only when there is sufficient information from that trial in order to calculate an estimate of the treatment difference and its standard error. In some cases, summary statistics which are already available from the trial can enable the calculations as if individual patient data were available. For example, in the case of a binary outcome, the knowledge of the number of success and failure is sufficient enough.

1.5: Multicentre trials and meta-analysis

Multicentre trials are usually conducted in order to enable the number of patients to be recruited within an acceptable period of time. Moreover, they multicentre trials provide a wider representation of the population compares to a single centre. Furthermore, that kind of trial will have been designed

prospectively with a combined analysis of the data from all centers as its main objective.

Individual centers have the same protocol, at least with respect to the collection of the data. When we have to do with a meta-analysis undertaken on a series of clinical trials, where we have a common outcome measure recorded and available individual data, then we can use the same linear modeling techniques as are applied to the analysis of a multicentre trial. In that case, “trial” will actually play the role of “center”. On the other hand, the analysis of a multicenter trial can be conducted using traditional meta-analysis methods, where “center” is playing the role of “trial”.

There is a continuum from the true multicenter trial, where all centers have an identical protocol, to a collection of trials addressing the same general therapeutic question but with different protocols. The known statistical methods can be used on that continuum, the choice and the validity of the results, however, of the most appropriate method can vary. Senn[21], in 2000, actually pointed out the differences between the traditional approaches applied to the analysis of multicenter trials and those on a meta-analysis. This probably have to do with the fact that most of the meta0analysis examines are retrospective and based on summary data from published papers.

https://doi.org/10.26219/heal.aueb.6071

CHAPTER 2

PROTOCOL DEVELOPMENT

2.1: Introduction

Before conducting a clinical trial it is a standard practice to prepare a study protocol to clearly specify the procedures to be followed. The same thought we must have for meta-analysis too, especially if with think that meta-analysis is even more complex process. In the case of an individual study it may be necessary to change the meta-analysis protocol due to unpredictable circumstances. Those changes must be reported and discussed in the results. Furthermore, in a meta-analysis protocol must be stated the key hypothesis of interest.

The protocol is a crucial procedure especially for a retrospective meta-analysis, or for one planned following the disclosure of the results from one or more trials. For such meta-analysis there is always the danger of bias due to study selection. There are cases where the meta-analysis can be performed only on a subset of the studies by virtue of inconsistency in the reporting or recording of outcome measures or even incompatible trial designs. Furthermore, if the meta-analysis is based only on published papers, the overall treatment difference will be estimated because usually only studies with statistically significant results are published.

Background information can help on setting a proper protocol for the meta-analysis, such as definition of the disease, possible incidents, and prognosis. Furthermore, the protocol must provide the reason for the conduction of the current meta-analysis. Moreover, the basic objectives of the meta-analysis should be stated clearly, in other words the study objectives.

2.2: Outcome Measures and baseline information

As we have already mentioned, each meta-analysis uses some outcome measures to perform the analysis. Those measures must be specified and analyzed in the protocol, providing operational definitions in case needed. Especially in case of an individual trial, it is recommended to specify the primary efficacy measures, so that the problem of multiple testing to be minimized.

Often, is important to have data of the baseline variables, such as demographic characteristics, prognostic factors and baseline assessments of efficacy and safety measures [5]. This kind of data can be useful in many cases. Firstly, they can be helpful in checking the comparability of patients who belong in different treatment categories, allowing such a way the within and between study comparison. Secondly, the availability of the individual patient data gives the opportunity to the analysts to conduct an analysis of covariance (ANCOVA) in which there have been made adjustments for one or more baseline variables in order to see if they have an important affect on the outcome measure. Finally, the baseline variables can be used to check the heterogeneity in the treatment difference across studies.

2.3: Source of data and study selection

In chapter 1, we referred to the publication bias. In order to minimize the problems associated with the selection bias, we must identify all the trials they could participate in the meta-analysis. This part of the protocol, must give information of the search strategy to be employed. When the meta-analysis is preplanned, there is no need of search strategy since all the trials, they are going to take part in our analysis, are identified before they are undertaken. For example, if a pharmaceutical company wants to conduct a retrospective meta-

https://doi.org/10.26219/heal.aueb.6071

analysis for its own drug is likely to know about all the relevant trials which have been made in the past. So, in that case there is no need for a search strategy.

However, in all other cases a search strategy is a crucial step. There is published information, including the online bibliography, databases of published and unpublished researches, dissertations and master's theses, trial registries etc. Moreover, in the protocol it must be specified the publication status and the time frame of the used bibliography. A forward search for example finds a publication and then searches all the later ones. On the contrary, a backward search identifies a publication and then finds all the earlier ones. Overall, the sources and the literature search must be a well-formulated, careful and co-ordinate effort involving several researches.

However, despite all the search and effort for finding as many information as we can, it is impossible to find all the relevant literature. In order to have a successful search, two aspects must be undertaken: recall and precision. We have the following definitions:

$$Recall = \frac{\text{Number of relevant documents retrieved}}{\text{Total number that should be retrieved}} \times 100$$

and

$$\text{Precision} = \frac{\text{Number retrieved and relevant}}{\text{Number retrieved}} \times 100$$

Recall measures the success of the retrieval process with a higher per cent corresponding to a more successful search [6]. However, the denominator is unknown, so it is not possible to know if the set of studies is representative. More advanced methods can be used, like an estimation of population size using capture-recapture models [8].

Precision has to do with the false positive rate. Our goal is to have both recall and precision literature retrievals high. It is possible to increase recall and precision if we use multiple methods of search strategies.

After collecting the source of data, we must clarify the selection criteria for studies in the meta-analysis. In case there is more than one hypothesis we want to test, it may be preferable to define separate criteria for each one. Furthermore, for each case we may have to create two separable groups of study. The first group is for the primary search we want to conduct, and the second one for the additional studies.

However, the selection of the criteria for the primary studies must be carefully defined. In case they are very strict, there is the possibility that the results will not be applicable in a vast population, but only in a subset of population or only for a very specific treatment. On the other hand, if they are too liberal there is a possibility to be impossible to be combined with the individual trial results.

In the report of the meta-analysis, it is necessary to include a list of all the studies that were excluded as well as the reason for that, along with the studies included. Sometimes, new information may come on the surface during the conduction of the meta-analysis. Consequently, that indicates a need to alter the study selection criteria.

2.4: Data extraction

In every meta-analysis, the data items they are going to be extracted must be specified. For that reason, it could be useful to produce an additional report which will have all the details for the desirable format for the data, the recommended coding and the data checking procedures.

In the case of a meta-analysis on individual patient data, since the results have nothing to do with the way in which the individual trial results are reported,

the information we gain is the most reliable. For such a meta-analysis, the goal should be to obtain individual patient data from all randomized subsets in all the included trials. This will give the opportunity to have a consistent approach towards the coding of data along with the handling of the missing data across all trials. If there is a common database structure among all trials then we have an easier task in integrating their data. However, for many meta-analyses, especially the retrospective ones, the data are not centrally located consequently more effort and time is required to collect all the necessary items together [7].

As we have mentioned in chapter 1, mane meta-analyses are conducted using only summary information from published papers. Even if the plan is to try to collect all the relevant individual patient data, in some cases that is not possible. In order to use those results, we must be careful and to evaluate the type and the kind of information which will be required. For example, in case we have a binary outcome, such as success or failure, we must choose a measure to calculate the treatment difference. Suppose that the measure we choose is the log-odds ratio of success on the new treatment versus the placebo. The number of successes and failures in each treatment is a sufficient knowledge. We can include a trial if the available data enable an estimation of the log-odds ratio and the calculation of its variance. But, if the knowledge of that data is the estimation of the difference in the success probabilities between the two treatment groups, then the trial cannot be included.

2.5: Statistical Analysis

The main protocol must include the basic features of the statistical analysis. Sometimes it is useful to have a separate statistical analysis plan, which will be more precise and more detailed. For each outcome variable, we must consider the following:

- The analysis population
- The missing data

- The analysis of individual trials
- The meta-analysis model
- The estimation and hypothesis testing
- The testing of heterogeneity and its exploration

Let's talk about those topics a little bit further.

2.5.1: Analysis population

The first step for a meta- analysis as we have already mentioned is to define the set of subjects who are going to take part in our study. Often, this will be based on the intention-to-treat principle. According to that principle, in respect of an individual trial, all randomized patients should be included in the analysis as members of the treatment group to which they were randomized [5]. This principle prevents the potential bias providing an objective basis for our statistical analysis.

In case where all randomized subjects satisfy all the selection criteria of the trial, provide complete data and are according to all trial procedures then the intention-to-treat is a straightforward implementation. Unfortunately, this ideal situation is difficult to be found in practice. If there is a reasonable explanation without the introduction of bias in the procedure, then there is the possibility to eliminate certain randomized subjects from the analysis set. According to ICH E9 we use the term full analysis "to describe the analysis set which is a complete as possible and as close as possible to the intention-to-treat ideal of including all randomized subjects".

It is an often tactic to report clinical trials which include analyses referring to a secondary subset of the total population, known as the "per protocol" set. The "per protocol" set contains patients who are more compliant with the protocol of the meta-analysis. For example, those patients are not classified as major protocol violators, they are able to provide data for the

primary analysis and they have completed a minimum period on study treatment. There are times when the analysis is undertaken to all subjects who complete the study period and are able to provide data for the primary efficacy variable. That is defined as a “completers” analysis and is an example of the “per protocol” analysis. However, due to relation between the treatment and the potential outcome, there is a possibility to have a biased analysis. For example, if we want to compare a new treatment over a placebo and we eliminate all the subjects who cannot tolerate the new treatment, then the analysis on the “per protocol” set will produce a large estimate of the treatment difference than the one based on the full analysis set.

Although most of the times a meta-analysis is taken place to see if one treatment is better than another one, there are times when a meta- analysis wants to determine of two treatments are equivalent. In the second case, since the conservative nature of the intention-to-treat approach may be inappropriate, the “per protocol’ set should be looked as a better choice for a meta-analysis than the “full analysis” set.

In case we want to conduct a meta-analysis using an individual patient data, it is more appropriate to use data from all randomized patients in order to achieve more completed and reliable results. Moreover, in the case when a meta-analysis is based on summary information from published papers or trial reports, problems may arise. Those problems have to do with the criteria of selection of the data that the authors used for their main analysis. Generally, there are papers which present results derived from a “full analysis” set and others using a “per protocol” set. It is advisable to separate in two groups the current results, one group with the result from a ‘full analysis” set and another one with the “per protocol” set before deciding if it is appropriate to combine them or not

.

2.5.2: Missing data

The most common problem in statistical analysis generally, is the handling of missing data. In the case of meta-analysis, the intention-to treat specifies the subjects who are going to take part in the analysis, but does not provide any solutions for the dealing of the missing data. Furthermore, in the case of an individual trial, the effect of the missing data on the overall results must to be mentioned.

In some cases, there are subjects who meet the criteria for the "full analysis" set but do not provide efficient are relevant information for some of the outcomes of interest, including sometimes and the primary efficacy variable. There are several reasons for that, it may happen because the subject got withdraw from the study so it does not provide further information from that point and on, or because it is lost to follow up.

There are several ways to deal with this situation. One option is to use for the analysis of each variable of interest only those subjects that provide data for those particular variables. It is important to understand that in that case the chosen set relies on the assumption that the reason for the current missing data is not the potential recorded values versus the actual ones, but the fact that data are missing at random [10]. In fact, if the mechanisms for dealing the missing data differ among the studies, then there is the fear of introducing bias into the estimate of the treatment difference.

Another way of dealing the missing data is to substitute values for the missing data. If the outcome of interest is measured in certain time points, then we can substitute the values that we do not have with earlier ones using imputation techniques. Imputation techniques range from carrying forward the last observation to the use of complex mathematical models [11], [12]. However, the analysts must be very careful when they use those techniques since there is a possibility that those techniques themselves may lead to biased estimates of the treatment difference. Furthermore, in some trials, data continue to be selected

even from subjects that have been eliminated during the study. Those subjects may use an alternative treatment after the elimination, so the analysts must be very careful.

If the amount of the missing data is large enough, then there is the possibility if questioning the analysis. When we perform a meta-analysis with individual patient data, then the methods of dealing with the missing data must be reported in the protocol. When a meta-analysis is based on summary information, the amount of missing data along with the methods that the authors used to deal with this problem must be under consideration in order to see the potential factors that can affect the meta-analysis.

2.5.3: Analysis of individual trials

It is important to present both the results from the individual trials along with the results from the meta-analysis. Often, the individual trial summaries may vary from the results from the earlier trial reports and publications due to the fact that it is preferable to take the same approach to the analysis of each trial in order to be consist with the meta-analysis. If we have the individual patient data, then a reanalysis is an easy task. However, that it is not possible when we have to deal with a meta-analysis based on summary information. In that case, we hope that the results will allow us to use the same measure of treatment difference in all studies.

Moreover, in the protocol we must specify the measure of treatment difference we will use in our analysis. For example, for binary data we may use the log-odds ratio. For a continuous data we might use the absolute difference in means.

2.5.4: The meta-analysis model

We must report the meta-analysis model we are going to use along with the specification of the terms which are going to be treated as fixed or random effects. When we perform a meta-analysis based on the assumption that all trials use the same parameter measuring treatment difference, then we typically refer to it as a "fixed-model" effect. On the other hand, we have a "random-effects" model when we allow this parameter to act as a random variable, taking different values from one trial to another.

2.5.5: The estimation and hypothesis testing

The basic hypothesis of the analysis must be specified. For example, when we want to compare a new treatment versus a standard one, the null hypothesis, which is no treatment difference, can be tested against the two-sided alternative of some difference between the two treatments. There are analyses where the treatment has been tested in more than one dose level. In that case, we should not combine data from all doses together, but to choose one dose level of primary interest. Furthermore, we can investigate the relationship between the dose and the response.

2.5.6: The testing of heterogeneity and its exploration

As we have already mentioned in chapter 1, most of the meta-analyses are retrospective on studies which they were not planned for a meta-analysis. The potential differences in the protocols of each study can produce heterogeneity. Moreover, we may have heterogeneity even when we have the same protocols among the studies, due to the quality difference of each study or the possible mistakes in implementing the protocol. Therefore, it is very

important to include a test of heterogeneity in the treatment difference parameter across studies.

Often, a test of heterogeneity is used to decide if we want to present an overall fixed effects estimate of difference or a random effects estimate. For example, in case we have a p-value equal or less than 0.05, then is better to calculate a random effects estimate or else we may calculate the fixed effects estimate. However, even if a statistical test of heterogeneity gives as useful descriptive information about the variability between trials, it is not recommended to use only the p-value for our decision.

The possible causes of heterogeneity can be identified in advanced, and the methods which we are going to be used to deal with them must be specified. Their investigation can be undertaken via the inclusion of covariate by treatment interaction terms in the meta-analysis model [5]. In case we find a statistically significant interaction, then it would be better to present the relationship between the magnitude of the treatment difference and the covariate. If we have a continuous variable, a graphical presentation of that relationship it could be very informative.

In the case when the covariate term represents a factor with few levels, then the treatment difference can be presented for each level, and we refer to it as a subgroup analysis. To test the hypothesis across all subgroups is the same as to test the hypothesis that the covariate of the treatment interaction is equal to zero.

2.6: Sensitivity Analysis

Extra thought must be given in performing sensitivity analyses in order to test the key assumptions which were made. Sometimes meta-analyses may be repeated if we exclude some trials. Additionally or alternatively, we can consider the results from some studies not classified as primary ones. The results can be

displayed along with the results of the primary studies in a graphical display. Furthermore, the meta-analyses can again be performed including these studies.

Another issue that must be taken into consideration is the systematic bias on the overall estimate of the treatment difference. When we are going to perform a retrospective analysis, or a meta-analysis conducted after some of the individual trials result are available, then the selection of the trials which are going to take part in the meta-analysis may introduce a systematic bias. In that case, we must estimate the possible impact of the bias in our analysis.

https://doi.org/10.26219/heal.aueb.6071

CHAPTER 3

ESTIMATING THE TREATMENT DIFFERENCE

3.1: Introduction

Many meta-analyses have to do with the comparison of two treatments in terms of a selected set of outcome measures. For each term separately, the goal is to estimate and to inference the potential difference between the effects of those two treatments. All this procedure involves the choice of an appropriate measure of the treatment difference, the calculation of the estimates of each study and an overall estimate of this difference. A traditional meta-analysis is one in which the overall estimate of treatment difference is calculated from a weighted average of the individual study estimates [5].

The traditional meta-analysis can be performed when the available data are in the form of study estimates, in studies based on summary statistics, in an individual patient data or in a combination of the different forms of study. The first step as we have already mentioned is to choose the appropriate measure to calculate the treatment difference. However, this is not always an easy task. Furthermore, it is a common phenomenon that the meta-analyst to have a little control over the selection of that measure. This is happening because most of the decisions are strongly related with the measures that were employed on the primary studies. For example, if in the primary studies it is used risk differences instead of survival times, then the analyst cannot do anything else but to use the average risk differences as the summary statistics.

Also, there is the possibility of different summary measures among the primary studies. Success may be defined as 30-day survival in one study, as a symptom-free in another. In that case, the analyst must create a comparable summary measure among the studies. Sometimes this is an impossible task, some

others not. In this chapter, we will describe three groups of measures outcome: measures based on discrete outcome data, measures based on continuous data and finally a miscellaneous set of outcome measures that are based on test statistics.

There are several ways to test the hypothesis concerning the treatment difference. In an individual trial the likelihood ratio test is often used to test the above hypothesis. The maximum likelihood estimate (ML) of the treatment difference is then presented along with the standard error or the confidence interval. The ML estimate is a widespread measure with the advantage of asymptotic optimality. Plus, it is available in every statistical package. Furthermore, the likelihood approach it can be extended from an individual trial to the meta-analysis as long as the individual patient data is available.

Another approach to perform the required estimations is based on the efficient score and Fisher's information statistics. This approach is very popular in the meta-analysis world and it is capable of producing an approximately ML estimates.

The notation we will use is the following:

θ: denotes the measure of treatment difference. It can be defined to be equal to zero when we have to equivalent treatments.

$\hat{\theta}$: denotes the estimate of θ

$var(\hat{\theta})$ and $se(\hat{\theta})$: the estimated variance and its standard deviation of $\hat{\theta}$ respectively

Z: the efficient score of θ under the null hypothesis that θ=0

V: the observed Fisher's information evaluated under the hypothesis that θ=0

T: the treatment group

C: the control group

When θ is small then we can use the distributional result Z~N(θV,V). Also, the estimate $\hat{\theta} = \frac{Z}{V}$ is an approximate ML estimate with variance equals to $\frac{1}{V}$ and

https://doi.org/10.26219/heal.aueb.6071

standard deviation equals to$\frac{1}{\sqrt{V}}$. The estimate of θ given by $\frac{Z}{V}$ is often called "one-step" estimate because it is obtained on the first step of Newton- Raphson procedure to maximize the log-likelihood function when the starting value for θ is 0. But, despite the fact that is asymptotically unbiased under the null hypothesis, it becomes bias as θ moves away from 0 [13].

3.2: Binary Data

A binary variable can take only two possible values, often "success" and "failure". To have a binary data we record the binary outcome for each patient. We denote with p_T the probability of success for a patient who belongs in the treated group and with p_C the probability of success for a patient from the control group. The treated group and the control group has n_T and $n_C$patients respectively. The number of successes and failures in the treated group are denoted by s_T andf_T. Accordingly, in the control group we have s_C successes and f_C failures. We can represent the data in a 2x2 table as shown in Table 3.1

Outcome	Treated Group	Control Group	Total
Success	s_T	s_C	s
Failure	f_T	f_C	f
TOTAL	n_T	n_C	n

Table 3.1: Data of parallel group study with a binary outcome

When we have a binary response variable, the knowledge of individual patient data gives nothing to the information shown in the Table 3.1 in order to estimate the treatment difference. Often, the summary statistics in a report or publication enables the construction of this table in order to calculate an identical estimate to that based on individual patient data.

https://doi.org/10.26219/heal.aueb.6071

For binary data, there are several measures of treatment difference that could be applied. Firstly, is the probability difference $p_T - p_C$. If the event we modeled is undesirable, for example the occurrence of a heart attack, the above difference is referred as risk difference. Secondly, is the log-odds ratio$\log\left[\frac{p_T(1-p_C)}{p_C(1-p_T)}\right]$. Thirdly, is the log-relative[2] risk$log\left(\frac{p_T}{p_C}\right)$.Table 3.2 shows the above measures of treatment.

	Probability Difference	**Log-odds Ratio**	**Log- relative Ratio**
Parameter	$p_T - p_C$	$log\left(\frac{p_T}{p_C}\right)$	$\log\left[\frac{p_T(1-p_C)}{p_C(1-p_T)}\right]$
Estimator	$\widehat{p_T} - \widehat{p_C}$	$\log\left(\frac{s_T f_C}{s_C f_T}\right)$	$\log\left(\frac{s_T/n_t}{s_C/n_C}\right)$
Standard Error	$\sqrt{\frac{s_T f_T}{n_T^3} + \frac{s_C f_C}{n_C^3}}$	$\sqrt{\frac{1}{s_T} + \frac{1}{s_C} + \frac{1}{f_T} + \frac{1}{f_C}}$	$\sqrt{\frac{f_T}{s_T n_T} + \frac{f_C}{s_C n_C}}$

Table 3.2 : Measures of treatment

From the above three measures, the most preferable is the log-odds ratio, because the relationship of the corresponding test statistics to their asymptotic normal or chi-squared distribution is closest [14]. The probability difference may have some problems because the accepted values for the difference are between -1 and +1, but the confidence intervals based on asymptotic theory allow points outside those limits. Another advantage of the log-odds ratio versus the log-relative risk is that if the probability of failure is put in place of the probability

[2] In that point we must mention that this name makes sense only when an undesirable event is being modeled.

of success, the resulting log-odds ratio will be of opposite sign and equal magnitude, whereas the log-relative risk will be of opposite sign but not of equal magnitude [5]. However, the main reason to use the log-odds ratio against to odds ratio is that the second one has a finite interval from 0 to 1 in order to represent the values corresponding to a lower relative success probability in the treated group and an infinite interval from 1 upwards for a higher relative success probability.

3.2.1: Log- odds ratio

We have the log-odds ratio

$$\theta = log\left\{\frac{p_T(1-p_C)}{p_C(1-p_T)}\right\} \tag{3.1}$$

The above logarithm is the log-odds of success on treatment relative to control[3]. The methods to analyze binary data with the use of full likelihood consider an unconditional distribution of the data based on binomial distribution. In that case, s_T and s_C are treated as observations from random variables.

The ML estimate can be found by fitting a linear logistic regression model. Furthermore, the ML estimate can be also calculated from the sample log-odds ratio, given by

$$\hat{\theta} = log\left(\frac{s_T f_C}{s_C f_T}\right) \tag{3.2}$$

The asymptotic estimate of variance, which is derived by the delta method [15] and used in the Wald test, is

$$\text{var}(\widehat{\theta}) = \frac{1}{s_T} + \frac{1}{s_C} + \frac{1}{f_T} + \frac{1}{f_C} \tag{3.3}$$

[3] The logarithm is to base e.

3.2.2: Probability difference

We have the probability difference

$$\theta = p_T - p_C \tag{3.4}$$

The unconditional ML estimate of that probability difference is calculated by the difference in the observed success probabilities as follows:

$$\hat{\theta} = \frac{s_T}{n_T} - \frac{s_C}{n_C} \tag{3.5}$$

The asymptotic estimate of variance is also derived using the delta method

$$\mathrm{var}(\hat{\theta}) = \frac{s_T f_T}{n_T^3} + \frac{s_C f_C}{n_C^3} \tag{3.6}$$

3.2.3: Log-relative risk

Consider setting the parameter θ equal to the log-relative risk

$$\theta = log\left(\frac{p_T}{p_C}\right) \tag{3.7}$$

The unconditional ML estimate is given by the sample log-relative risk

$$\hat{\theta} = \log\left(\frac{s_T/n_t}{s_C/n_C}\right) \tag{3.8}$$

The asymptotic estimate of variance is the following:

$$\mathrm{var}(\hat{\theta}) = \frac{f_T}{s_T n_T} + \frac{f_C}{s_C n_C} \tag{3.9}$$

3.2.4: Example

We will illustrate an example of a meta-analysis with binary data, first presented by Collins et al. in 1990. This meta-analysis includes 14 randomized trials of antihypertensive drugs. These trials were conducted in patients with hypertension comparing an antihypertensive treatment versus a placebo or "usual care". The trials were separated into groups according to the level of hypertension of the patients. We have four trials which include patients with mild hypertension (i.e. diastolic blood pressure (DBP)< 110 mmHg) and three trials with patients with mild to tolerate hypertension (DBP< 115 mmHg). One of the trial, the Hypertension Detection and Follow-up Program study, presented the results in such a way that patients with DBP< 110 mmHg and 110-115 mmHg were examined separately from those who had DBP>115 mmHg. In our example, each category of patients will be examined separately. Our interest has to do with the effects of anti-hypertension treatment on stroke. Patients were monitored for an average of 5 years. The following table presents the number of patients who had a stroke in each treatment.

Study	Treated Group		Control Group		Treated group % stokes	Control group % strokes
	Number of strokes	Total number	Number of strokes	Total number		
Trials in which all patients had entry DBP<110 mmHG						
1 VA- NHLB1	0	508	0	504	0,0	0,0
2 HDFP (Stratum I)	59	3903	88	3922	1,5	2,2
Oslo	0	406	5	379	0,0	1,3
ANBPS	13	1721	22	1706	0,8	1,3
5 MRC	60	8700	109	8654	0,7	1,3
Trials in which all patients had entry DBP?115 mmHG						
6 VAII	5	186	20	194	2,7	10,3
7 USPHS	1	193	6	196	0,5	3,1
8 HDFP (Stratum II)	25	1048	36	1004	2,4	3,6
9 HSCSG	43	233	52	219	18,5	23,7
Trials in which some or all patients had entry DBP>115 mmHG						
10 VAI	1	68	3	63	1,5	4,8
11 WOLFF	2	45	1	42	4,4	2,4
12 Barraclough	0	58	0	58	0,0	0,0
13 CartER	10	49	21	48	20,4	43,8
14 HDFP (Stratum III)	18	534	34	529	3,4	6,4
15 EWPHE	32	416	48	424	7,7	11,3
16 Coope	20	419	39	465	4,8	8,4
TOTAL	289	18487	484	18407	1,6	2,6

Table 3.3: The number of hypertensive experiencing a stoke

Firstly, we will construct the table showing the number of strokes in patients in study 2 in the format of table 3.1:

Outcome	Treated Group	Control Group	Total
Success	59	88	147
Failure	3844	3834	7678
Total	3903	3922	7825

Table 3.4: Number of patients for study 2

To calculate the estimate of θ and the variance of the estimate of θ with the log-odds theory we mentioned above, we will use the formulae (3.2) and (3.3).

https://doi.org/10.26219/heal.aueb.6071

$$\hat{\theta} = log\left(\frac{59 \times 3834}{88 \times 3844}\right) = -0.402$$

and

$$\text{var}(\widehat{\theta}) = \frac{1}{59} + \frac{1}{88} + \frac{1}{3844} + \frac{1}{3834} = 0.029$$

For the probability difference, we use the formulae (3.5) and (3.6)

$$\hat{\theta} = \frac{59}{3903} - \frac{88}{3922} = -0.00732$$

and

$$\text{var}(\hat{\theta}) = \frac{59 \times 3844}{3909^3} + \frac{s_C f_C}{3922^3} = 0.0000094$$

Finally, for the log-relative risk we use the types (3.8) and (3.9) to calculate the estimate and variance of a stroke on antihypertensive treatment compared with control in study 2

$$\hat{\theta} = \log\left(\frac{59/3903}{88/3922}\right) = -0.395$$

and

$$\text{var}(\hat{\theta}) = \frac{3844}{59 \times 3903} + \frac{3834}{88 \times 3922} = 0.028$$

3.3: Normally distributed Data

When we have to deal with quantitative measurement on a continuous scale, we often treat it as a normally distributed data. Even if it is not possible at once, we can transform the values in such a way in order to produce a normal distribution data. The subjects that belong in the treated group are modeled as being normally distributed with mean μ_T and standard deviation σ. The subjects in the control group have mean μ_T and standard deviation σ. We have the assumption that there is a common standard deviation between-patient within each treatment group.

We have n_T subjects in the treated group with responses y_{jT} with j=1, 2,…..,n_T and n_C subjects in the control group with responses y_{jC}, with j=1,2,…., n_C. For the treated group we have the following:

- the sample mean: $\overline{y_T} = \frac{1}{n_T}\sum_{j=1}^{n_T} y_{jT}$
- sample standard deviation: $s_T^2 = \frac{\left(\sum_{j=1}^{n_T} y_{jT}^2\right) - n_T\overline{y_T^2}}{n_T - 1}$
- the sum of the observations $A_T = \sum_{j=1}^{n_T} y_{jT}$
- and the squares of the observations $B_T = \sum_{j=1}^{n_T} y_{jT}^2$

With similar way, we define $\overline{y_C}$, s_C^2, A_C and B_C for the control group. Furthermore, we have that

$A = A_T + A_C$ and $B = B_C + B_T$. We summarize data in the following table 3.5:

Data	Treated group	Control group	Total
Number of patients	n_T	n_C	n
Mean	$\overline{y_T}$	$\overline{y_C}$	
Standard deviation	s_T^2	s_C^2	
Sum of observations	A_T	A_C	A
Sum of squares of observations	B_T	B_C	B

Table3. 5: Data for the parallel group study with normally distributed outcomes

The published reports present the above table with the number of patients, the sample mean and sample standard deviation for each treatment group. Moreover, that is the only information we need in order to calculate the treatment difference. The values A_T and B_T are calculated using the sample mean and the sample standard deviation as follows:

$$A_T = n_T\overline{y_T^2} \tag{3.10}$$

and

$$B_T = (n_T - 1)s_T^2 + n_T\overline{y_T^2} \tag{3.11}$$

For the data which is normally distributed there are two parameters of treatment difference which are considered: the absolute difference between means and the standardized difference between means. The absolute difference $\mu_T - \mu_C$ is easy enough to be interpreted plus it is more appropriate measure if all the studies have the same measurement. However, the standardized difference$\frac{\mu_T - \mu_C}{\sigma}$, due to the fact that it is dimensionless, it can be used when we want to combine different units or scales. Furthermore, it gives the possibility to

calculate the probability that a patient in the treated group will do better than a patient in the control group.

3.3.1: Absolute different between means

The ML estimate of the absolute difference between means, using the full likelihood, is the difference between the sample means:

$$\hat{\theta} = \bar{y}_T - \bar{y}_C \tag{3.12}$$

The variance is:

$$\text{var}(\widehat{\theta}) = \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \tag{3.13}$$

However, for the calculation of the above variance we must firstly choose an appropriate estimate forσ^2. One choice is the ML estimate $\hat{\sigma}_M^2$, where:

$$\hat{\sigma}_M^2 = \frac{B_T - {A_T^2}/{n_T} + B_C - {A_C^2}/{n_C}}{n} \tag{3.14}$$

However, since it is known that formula (3.14) is biased, there is a more common estimate for theσ, which is the sample standard deviations^2. $s^2$is an unbiased estimate which is obtained by replacing the dominator in the formula (3.14) by n-2, i.e.

$$s^2 = \frac{B_T - {A_T^2}/{n_T} + B_C - {A_C^2}/{n_C}}{n-2} \tag{3.15}$$

The ML estimate $\hat{\theta}$ as long as the variance using the s^2 is easy to be found using statistical packages, since we can simply fit a general linear model.

3.3.2: Standardized difference between means

The ML estimate of the standardized difference between means is:

$$\hat{\theta} = \frac{\bar{y}_T - \bar{y}_C}{\hat{\sigma}_M} \tag{3.16}$$

Often, in the formula (3.16) we use the unbiased estimate ofσ^2:

$$\hat{\theta} = \frac{\bar{y}_T - \bar{y}_C}{s} \tag{3.17}$$

In both cases, the approximate variance is calculated:

$$\mathrm{var}(\hat{\theta}) = \frac{n}{n_T n_C} \tag{3.18}$$

In 1976, Glass [15] proposed an estimate of σ calculated only from the control group. When we compare several treatments with the control group, the pairwise comparisons of each treated group with the control one could lead to different standardized values of identical mean differences. Despite the fact that the sample standard deviations will most of the times differ among the treatment groups, in many cases we do the assumption of a common variance.

Hedges and Olkin[16] proved that the estimate of θ given by the formula (3.17) is biased and the defined a new estimate of θ in order to remove the mentioned bias:

$$\hat{\theta} = \frac{J(n-2)(\bar{y}_T - \bar{y}_C)}{s} \tag{3.19}$$

With variance estimate:

$$\mathrm{var}(\hat{\theta}) = \frac{n}{n_T n_C} + \frac{\hat{\theta}^2}{2n} \tag{3.20}$$

where J(m) is listed in Table 3.6 and which for large m can be calculated:

https://doi.org/10.26219/heal.aueb.6071

$$J(m) \approx 1 - \frac{3}{4m - 1} \tag{3.21}$$

As m getting bigger and bigger, J(m) approaches unity, as a result the distributions of the estimates in (3.17) and (3.19) tend to be normally distributed with identical means and variances.

m	J(m)	m	J(m)	m	J(m)	m	J(m)
2	0.5642	15	0.949	28	0.9729	41	0.9816
3	0.7236	16	0.9523	29	0.9739	42	0.982
4	0.7979	17	0.9551	30	0.9748	43	0.9824
5	0.8408	18	0.9577	31	0.9756	44	0.9828
6	0.8686	19	0.9599	32	0.9764	45	0.9832
7	0.8882	20	0.9619	33	0.9771	46	0.9836
8	0.9027	21	0.9638	34	0.9778	47	0.9839
9	0.9139	22	0.9655	35	0.9784	48	0.9843
10	0.9228	23	0.967	36	0.979	49	0.9846
11	0.93	24	0.9684	37	0.9796	50	0.9849
12	0.9359	25	0.9699	38	0.9801		
13	0.941	26	0.9708	39	0.9806		
14	0.9453	27	0.9719	40	0.9811		

Table 3.6: Values for the factor J(m)

3.3.3: Example

In this section, we will illustrate an example [5]. We have a multicenter study in order to compare two anaesthetic agents, A and B, in patients who have been undergone in small surgical procedures where a rapid recovery is crucial. Here, we have data from nine centers, which are considered to be as separate studies which are included in the meta-analysis. The response of interest is the recovery time, which is defined as the time when the anaesthetic gases were turned off until the time where the patient opens their eyes. The time is calculated in minutes. After a logarithmic transformation to the data, we consider the data to be normally distributed. The data is represented in the following table:

Center	Treatment A			Treatment B		
	Number of patients	Mean	Standard Deviation	Number of patients	Mean	Standard Deviation
1	4	1,141	0,967	5	0,277	0,62
2	10	2,165	0,269	10	1,519	0,913
3	17	1,79	0,795	17	1,518	0,849
4	8	2,105	0,387	9	1,189	1,061
5	7	1,324	0,47	10	0,456	0,619
6	11	2,369	0,401	10	1,55	0,558
7	10	1,074	0,67	12	0,265	0,502
8	5	2,583	0,409	4	1,37	0,934
9	14	1,844	0,848	19	2,118	0,749

Table 3.7: Recovery time after anaesthesia which is logarithmic transformed.

In the following table we give the results for center 1 of the study following the format of Table 3.5.

Data	Treated group	Control group	Total
Number of patients	4	5	9
Mean	1.141	0,28	
Standard deviation	0,97	0,62	
Sum of observations	4.564	1,39	5.949
Sum of squares of observations	8.013	1,92	9.934

For the absolute difference between means we have:

$$\hat{\theta} = 0.864$$

$$s^2 = 0.621$$

and

$$\text{var}(\hat{\theta}) = 0.279$$

For the standardized difference between means we have the following results:

$$\hat{\theta} = \frac{0.864}{0.788} = 1.097$$

https://doi.org/10.26219/heal.aueb.6071

and

$$\text{var}(\hat{\theta}) = 0.450$$

Using the formulas (3.17) and (3.19) we have

$$\hat{\theta} = 0.973 \text{ and } \text{var}(\hat{\theta}) = 0.503$$

3.4: Ordinal data

We used the Clinical Global Impression of Change (CGIC) scale to produce an overview by the clinician of whether a patient with Alzheimer's disease is getting better or worse. It is a seven-point scale that is intended to assess change from baseline, where scores 1,2 and 3 denote "very much improved", "much improved" and "minimally improved" respectively, 4 shows "no change" and 5,6 and 7 represent "minimally worse", "much worse" and "very much worse". We illustrate the results from five trials comparing tacrine with placebo, in the following table 3.8, in which the CGIC scale was used. Qizilbash et al (1998) represented a meta- analysis using these data[23].

Study	Treatment	CGIC scale							Total
		1	2	3	4	5	6	7	
1	Tacrine	2	2	23	45	22	2	0	96
	Placebo	0	2	22	54	29	3	0	110
2	Tacrine	0	14	119	180	54	6	0	373
	Placebo	0	1	22	35	11	3	0	72
3	Tacrine	1	12	20	24	10	1	0	68
	Placebo	0	7	16	17	10	3	0	53
4	Tacrine	3	18	106	175	62	15	2	381
	Placebo	0	8	24	73	52	13	0	170
5	Tacrine	0	3	14	19	3	0	0	39
	Placebo	0	2	13	18	7	1	0	41

Table 3. 8: Number of patients in each category of the CGIC scale in the tacrine studies

https://doi.org/10.26219/heal.aueb.6071

As we can see from the above table, most of the patients were placed in the middle three categories, with hardly any in the two extreme categories. In this example, we will combine the first two categories, 1 and 2, as long as the categories 6 and 7, so as to have a five-category response.

3.4.1: Measurement of treatment difference

The patients responses fall into one of the m categories, $C_1,\ldots\ldots, C_m$, which are ordered with C_1 as the best and C_m as the worst response. The mathematical model used here is expressed in terms of the probability of falling into category k, given by p_{kT} and p_{kC} with k=1, …, m for the treated and the control group respectively. Cumulative probabilities of falling into category C_k for the treated and control group are denoted by Q_{KT} and Q_{KC} respectively, with:

$$Q_{KT}=p_{1T}+\ldots+p_{kT} \quad \text{and } Q_{KC}=p_{1C}+\ldots+p_{kC}$$

The data can be presented in the form of mx2 table, as the one below.

Number of patients in category	Treated group	Control group	Total
C_1	n_{1T}	n_{1c}	n_1
.	.	.	.
.	.	.	.
.	.	.	.
Cm	n_{mT}	n_{mC}	n_m
Total	n_T	n_C	n

Table 3.9: Data for a parallel group study with an ordinal outcome

This table provides the entire summary we need to estimate the treatment difference. Suppose we are in position to create such a table with the

necessary information, the estimate will be identical to that based on individual patient data.

For ordinal data, we will present two measures of treatment difference, the log-odds ratio based on the proportional odds model and the log-odds ratio based on the continuation ratio model.

3.4.2: Log-odds ratio (proportional odds model)

Consider the following log-odds ratio:

$$\theta = log\left\{\frac{Q_{kT}(1-Q_{kC})}{Q_{kC}(1-Q_{kT})}\right\} \qquad (3.22)$$

We make the assumption that θ is constant over all values of k. The parameter θ can be treated in the same way as the log- odds ratio for binary data, i.e. the log-odds of being better off on treatment relative to control. If we reduce the data in two levels outcome, i.e. success/failure, with categories $C_{1,}\ldots C_k$ representing success and $C_{k+1},\ldots,C_m$ representing failure. In that case, Q_{kT} and $1\text{-}Q_{kT}$ are the probabilities of success and failure respectively for the treated group, and Q_{kC} and $1\text{-}Q_{kC}$ are similarly defined for the control group. For the m categories we have m-1 splits. The proportional odds assumption is equivalent to supposing that all m-1 binary analyses refer to the same log-odds ratio θ. In the case that we have only two response categories, as we mentioned before success/failure, then the proportional odds model is the same as the usual linear logistic model for binary data.

Consider the table 3.10, which is a 5x2 table created for study 1. The ML estimate and the variation for this study are given by:

$$\hat{\theta} = 0.284 \qquad \text{and} \qquad \text{var}(\hat{\theta}) = 0.068$$

Treatment	Category					Total
	C1	C2	C3	C4	C5	
Tacrine	4	23	45	22	2	96
Placebo	2	22	54	29	3	110

Table 3. 10: CGIC data from tacrine study 1

If we use a marginal likelihood based on the ranks with allowance for ties, then the test statistics Z and V are the below[24]:

$$Z = \frac{1}{n+1}\sum_{k=1}^{m} n_{kC}(L_{kT} - U_{kT}) \tag{3.23}$$

and

$$V = \frac{Z^2}{n+2} + \frac{B}{(n+1)(n+2)} \tag{3.24}$$

where

$$L_{kT} = n_{1T} + \cdots . + n_{(k-1)T} \qquad k = 2, \ldots m$$

$$U_{kT} = n_{(k+1)T} + \cdots . + n_{mT} \qquad k = 1, \ldots m-1$$

$$L_{1T} = U_{mT} = 0$$

$$B = \sum_{k=1}^{m} \{n_{kT}(n_C - n_{kC}) + n_{kT}n_{kC}(n - n_k) + 2n_{kT}L_{kC}U_{kC} + 2n_{kC}L_{kT}U_{kT}\} \tag{3.25}$$

Similarly, we have the definitions for L_{kC} and U_{kC}.

Moreover, an approximate large- sample formula for V is

$$V' = \frac{n_T n_C n}{3(n+1)^2}\left\{1 - \sum_{k=1}^{m}\left(\frac{n_k}{n}\right)^3\right\} \tag{3.26}$$

Using the data from the table 3.8, we calculate the estimates:

$$Z = 4.155 \quad , \quad V = 14.668$$

$$\hat{\theta} = \frac{Z}{V} = 0.283 \quad \text{and} \quad \text{var}(\hat{\theta}) = \frac{1}{V} = 0.068$$

Formula (3.26) gives V'=14.611, very close to V.

A further interpretation of θ can be made in case where the categorization is made as a result of grouping originally continuous data. Firstly, a continuous response Y is observed and then the categorization follows as category C_k, if Y lies between α_{k-1} and α_k, for an increasing sequence of numbers $\alpha_0, \ldots, \alpha_m$. In the case that the response of a subject in the treated and in control group is denoted by Y_T and Y_C respectively, then the probability that the person in the treated group does better is:

$$P(Y_T > Y_C) = \frac{1 - e^{-\theta} - \theta e^{-\theta}}{(1 - e^{-\theta})^2}$$

3.4.3: Log- odds ratio (continuation ratio model)

Here, we have the probability that a subject belongs to a specific category conditional on being in that category or a worse one. This is a sort of "discrete hazard", but it is on favor of a desirable outcome. We define:

$$h_{kT} = \frac{p_{kT}}{1 - Q_{(k-1)T}} \tag{3.27}$$

where $Q_{0T=0}$. In a similar way we have the definition for h_{kC}.

Now, consider the log-odds ratio:

$$\theta = log\left\{\frac{h_{kT}(1 - h_{kC})}{h_{kC}(1 - h_{kT})}\right\} \quad , \text{for } k = 1, \ldots m - 1 \tag{3.28}$$

Due to the fact that the hazard is of a desirable event, we assume that θ is positive if the treated group is better than the control one.

https://doi.org/10.26219/heal.aueb.6071

It is obvious that for k=1, that both the log-odds ratio based on the proportional odds model or based on the continuation ratio are the same. Due to that, the estimates from the two approaches are likely to be of the same order of magnitude. McCullagh named the proportional odds model "palindromic invariant", meaning that if we model the cumulative probabilities starting from the best to worst case will only change the sign and not the magnitude of the parameter estimates, which would be obtained by starting from the worst to the best case. However, when we have a continuation model there is no "palindromic invariant". In that model we must decide from the beginning whether to model the hazard of a desirable or an undesirable event.

In the case we chose the continuation ratio model, the estimation of the corresponding log-odds ratio proceeds as follows. We use the data from the table 3.11:

	Treated group	Control group	Total
Number of patients in categories C_1 to C_{m-1}	$L_{(m-1)T}$	$L_{(m-1)C}$	$L_{(m-1)}$
Number of patients in categories			
C_1	n_{1T}	n_{1C}	n_1
.	.	.	.
.	.	.	.
C_{m-1}	$n_{(m-1)T}$	$n_{(m-1)C}$	$n_{(m-1)}$
Number of patients in the same or a worse(higher)category than			
C_1	R_{1T}	R_{1C}	R_1
.	.	.	.
.	.	.	.
C_{m-1}	$R_{(m-1)T}$	$R_{(m-1)C}$	$R_{(m-1)}$

Table 3. 11: Data for a parallel group study with ordinal data.

https://doi.org/10.26219/heal.aueb.6071

The likelihood which conditions on the rank R_k, k=1,…,m-1, is equivalent to the discrete form of the parallel likelihood under the Cox proportional hazards model. Therefore, it is possible to calculate the ML estimate for the log-odds ratio from any statistical package which fits the proportional hazards model. To continue the analysis, a response variable, denoted by cat, would be calculates to take the value k if the patient had a response to category k. This particular response will then be treated as a survival tome, with values in the highest category considered as censored observations with value m[5].

The following table 3.12 shows the data from tacrine study 1:

Category	Tacrine			Placebo		
	Number in worse category	Number in category	Number in this or worse category	Number in worse category	Number in category	Number in this or worse category
1	92	4	96	108	2	110
2	69	23	92	86	22	108
3	24	45	69	32	54	86
4	2	22	24	3	29	32

Table 3. 12: Data from study 1.

The ML estimate and its variance are:

$$\hat{\theta} = 0.227 \quad \text{and} \quad \text{var}(\hat{\theta}) = 0.050$$

For the calculation of the values of the efficient score and Fisher's information Z and V we have the formulae:

$$Z = L_{(m-1)T} - \sum_{k=1}^{m-1} \frac{n_k R_{kT}}{R_k} \tag{3.29}$$

$$V = \sum_{k=1}^{m-1} \frac{n_k (R_k - n_k) R_{kT} R_{kC}}{(R_k - 1) {R_k}^2} \tag{3.30}$$

The statistic Z is the log-rank statistic and the associated score test is the log-rank test. For the study 1, using the above formulae, we have:

$$Z = 4.576\,, \quad V = 20.190$$

$$\hat{\theta} = \frac{Z}{V} = 0.227 \quad \text{and} \quad \text{var}(\hat{\theta}) = \frac{1}{V} = 0.050$$

An alternative to the conditional likelihood approach we described above is a full likelihood, based on the multinomial distribution. In this approach, we have that each patient contributes multiple recordings of binary data, which depends on the category into which their response falls. If the patient's response falls into category k, then it contributes with success in that category and failure in the categories 1 to k-1. The categories which are greater than k, have no contribution. The likelihood can be presented in terms of the conditional probabilities h_{kT} and h_{kC}, for k=1,…m-1. The relation of these binary data with the log-odds ratio θ, is given by a binary model with a logit link function, i.e.

$$\log\left\{\frac{h_{kT}}{(1-h_{kT})}\right\} = \alpha_k + \theta \qquad and \qquad \log\left\{\frac{h_{kC}}{(1-h_{kC})}\right\} = \alpha_k \qquad (3.31)$$

where α_k is the parameter associated with the kth category.

In that case, the ML estimate can be calculated through any statistical package, using a logit link function. A patient with a response in category k provides a binary outcome, denoted resp, for categories 1,.., k. Furthermore, for the categories 1,..k-1 we have resp=0 and for category k we have resp=1.

The estimate for the study 1 and its variance are:

$$\hat{\theta} = 0.228 \quad \text{and} \quad \text{var}(\hat{\theta}) = 0.050$$

In the case we have a large number of categories the above approach is unsatisfactory plus with the fact that a model cannot be fitted due to the fact that if we have zero cells in the mx2 table there is the possibility to be lead in non-convergence. In such cases, the best approach is the one based on the conditional likelihood.

The Z and V statistics based on the full likelihood are:

https://doi.org/10.26219/heal.aueb.6071

$$Z = L_{(m-1)T} - \sum_{k=1}^{m-1} \frac{n_k R_{kT}}{R_k} \tag{3.32}$$

$$V = \sum_{k=1}^{m-1} \frac{n_k (R_k - n_k) R_{kT} R_{kC}}{{R_k}^3} \tag{3.33}$$

If we compare the Z and V from the conditional and the unconditional approach, we can see that V differs by a factor $(R_k - 1)/R_k$ in the denominator of each summand, when Z remains the same.

https://doi.org/10.26219/heal.aueb.6071

CHAPTER 4

COMBINING ESTIMATES OF A TREATMENT DIFFERENCE

4.1: Introduction

In this chapter we will see the methodology used for combining the study estimates of a treatment difference. This methodology can be firstly used when the data from each study contains only estimates of treatment difference along with their standard deviation or summary statistics. Also, it can be used when the individual patient data is available although in that case is better to use even more advanced statistical methods for better results. Furthermore, the methodology we will present in this chapter is appropriate to be used to combine studies some of which have available the individual patient studies with others providing only estimates of treatment difference or summary statistics.

We do have assumptions to combine estimates of treatment difference, as like we had in the previous chapters. First of all, there is a parallel group design comparing a treated with control groups. Furthermore, we assume that we will conduct a meta-analyses based on the same outcome measure which have been recorded in the same way in each trial along with the same parameterization of the treatment difference.

In this chapter, we will also present a general fixed effects parametric approach, which can be applicable in many different data types. This also includes two things: a calculation of an overall estimate of treatment difference and a statistical testing of the null hypothesis, i.e. that there is not a difference between the two treatments. Moreover, we will represent a statistic which tests the appearance or not of heterogeneity between the study parameters of treatment difference.

When we have a retrospective meta-analysis, it is possible to find differences in the study protocols, which is likely to lead in heterogeneity. In other situations, we may have evidence of heterogeneity from the study estimates. However, if we do have heterogeneity, it is not enough just to mention it. We must explore the reasons for its appearance.

4.2: A general fixed effects parametric approach

4.2.1: A fixed effects meta-analysis model

Here, we suppose that we have *r* independent studies each comparing the treated with the controlled group. Also, we have the same outcome measure in every study reported for each patient. We denote with θ the parameter which represents the measure of treatment difference. For example, in the case of normally distributed data this parameter may be the difference between the treatment means or in the case of binary data it could be the log-odds ratio. When the two treatments have an equal effect, we assume that θ equals 0. We denote with $\widehat{\theta}_{\mathrm{i}}$the estimate of θfor the *i*-th study. Then the general fixed effects model is given by:

$$\hat{\theta}_{\mathrm{i}} = \theta + \varepsilon_i \tag{4.1}$$

For i=1,2….r. The $\varepsilon_i$are the errors and are realizations of normally distributed random variables with expected values 0 and variance denoted by ξ_{i}^2. We have that

$$\hat{\theta}_{\mathrm{i}} \sim \mathrm{N}(\theta, \xi_{\mathrm{i}}^2) \tag{4.2}$$

4.2.2: Estimation and hypothesis testing of the treatment difference

Often, the variance of $\hat{\theta}_i$, var($\hat{\theta}_i$) is treated as it was the true variance ξ_i^2 so, there is no place for the calculation of an error in the term var($\hat{\theta}_i$). We denote with w_i the estimated inverse variance of $\hat{\theta}_i$, i.e. $w_i = \frac{1}{\text{var}(\hat{\theta}_i)}$. We make the assumption that the distribution is

$$\hat{\theta}_i \sim N(\theta, w_i^{-1}) \tag{4.3}$$

for i=1,.., r. Under the null hypothesis that the treatment difference in each study equals to 0, we have

$$\hat{\theta}_i w_i \sim N(0, w_i) \tag{4.4}$$

for i=1,.. r. Due to the fact that the study estimates are independent, we have:

$$\sum_{i=1}^{r} \hat{\theta}_i w_i \sim N\left(0, \sum_{i=1}^{r} w_i\right) \tag{4.5}$$

To compare the global null hypothesis that the treatment difference equals to 0 studies, is tested by comparing the statistic:

$$U = \frac{\left(\sum_{i=1}^{r} \hat{\theta} w_i\right)^2}{\sum_{i=1}^{r} w_i} \tag{4.6}$$

following the chi-square distribution with one degree of freedom. We assume that the studies have a common treatment difference, we have:

$$\sum_{i=1}^{r} \hat{\theta} w_i \sim N\left(\theta \sum_{i=1}^{r} w_i, \sum_{i=1}^{r} w_i\right) \tag{4.7}$$

The fixed treatment effect can be estimated using $\hat{\theta}$:

$$\hat{\theta} = \frac{\sum_{i=1}^{r} \hat{\theta}_i w_i}{\sum_{i=1}^{r} w_i} \tag{4.8}$$

If $w_i$were the true inverse variance and not an estimate, then $\hat{\theta}$ would be the maximum likelihood estimate of θ. The standard error of $\hat{\theta}$ is:

$$\mathrm{se}(\hat{\theta}) = \sqrt{\frac{1}{\sum_{i=1}^{r} w_i}} \tag{4.9}$$

and the confidence interval (CI) for θ is given by:

$$\hat{\theta} \pm 1.96\sqrt{\frac{1}{\sum_{i=1}^{r} w_i}} \tag{4.10}$$

In order to calculate all the above, we need an estimate of the treatment difference and its variance from each study. Usually, every trial report provides the reader with the standard error, so the calculation of w_i can be$\frac{1}{\{\mathrm{se}(\hat{\theta}_i)\}^2}$. In case of using efficient score and Fisher's information statistics, then $\hat{\theta} = \frac{Z_i{}^2}{V_i}$. For the choice of $\hat{\theta}_i$ it follows that $w_i = V_i$. Furthermore, $\hat{\theta}_i w_i = Z_i$and $\hat{\theta}_i^2 w_i = \frac{Z_i^2}{V_i}$. So, we have:

$$\hat{\theta} = \frac{\sum_{i=1}^{r} Z_i}{\sum_{i=1}^{r} V_i} \tag{4.11}$$

Plus,

$$U = \frac{(\sum_{i=1}^{r} Z_i)^2}{\sum_{i=1}^{r} V_i} \tag{4.12}$$

Sometimes, we refer to the fixed effects model as an "assumption free" approach because it is argued that the estimate of the treatment difference does not depend on the assumption of a common treatment difference parameter across all the including studies. If we consider that we have relaxed the assumption that we must have a common treatment difference in all studies, plus that the distribution of the individual study estimates becomes

$$\hat{\theta}_i \sim N(\theta_i, w_i^{-1}) \tag{4.13}$$

where θ_i is the treatment difference parameter in study i, then the overall fixed effect estimate $\hat{\theta}$ can now be calculated as an estimate of

$$\frac{\sum_{i=1}^{r} \theta_i w_i}{\sum_{i=1}^{r} w_i} \tag{4.14}$$

the weighted mean of the study treatment difference parameter. However, despite the fact that is an acceptable interpretation of $\hat{\theta}$, once the variation between studies is calculated, it would seem necessary to investigate the amount of heterogeneity.

4.2.3: Testing the heterogeneity across all studies

In order to calculate the heterogeneity in the treatment difference across the studies, we must use a large amount of sample. The above comment is based on the statistic:

$$Q = \sum_{i=1}^{r} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2 \tag{4.15}$$

which is the weighted sum of squares of the deviations of individual study estimates from the overall estimate[17]. If the treatment difference parameters are homogeneous, then Q follows the chi-square distribution with r-1 degrees of freedom. In that case, an easier but equivalent formula for calculation is:

$$Q = \sum_{i=1}^{r} \hat{\theta}_i^2 w_i - U \tag{4.16}$$

Moreover, when we use the efficient score and Fisher's information statistics, then Q can be written as:

$$Q = \sum_{i=1}^{r} V_i \left(\frac{Z_i}{V_i} - \frac{\sum_{i=1}^{r} Z_i}{\sum_{i=1}^{r} V_i} \right)^2 = \sum_{i=1}^{r} \left(\frac{Z_i^2}{V_i} \right) - \frac{(\sum_{i=1}^{r} Z_i)^2}{\sum_{i=1}^{r} V_i} \quad (4.17)$$

4.3: A general random approach effects parametric approach

4.3.1: A random effects meta-analysis model

As we have mentioned in earlier chapters, in a random effect model we assume that the treatment parameters in the r studies $(\theta_1, \ldots\ldots, \theta_r)$ are a sample of individual observations from $N(\theta, r^2)$. The general random effects model is given by the formula:

$$\hat{\theta}_i = \theta + v_i + \varepsilon_i \quad (4.18)$$

For i=1,…,r. We denote with v_is the normally distributed random effects with mean 0 and variance τ^2. Furthermore, we assume that both v_i and ε_i are independently distributed. It follows that

$$\hat{\theta}_i \sim N(\theta, \xi_i^{-1} + \hat{\tau}^2) \quad (4.19)$$

4.3.2: Estimation and hypothesis testing of the treatment difference

In most cases, we do not know the parameter τ^2 and we must estimate it. Therefore, the distributional assumption we make is that:

$$\hat{\theta}_i \sim N(\theta, w_i^{-1} + \hat{\tau}^2) \quad (4.20)$$

where $\hat{\tau}^2$ is the estimation of τ^2. If we set

$$w_i^* = (w_i^{-1} + \hat{\tau}^2)^{-1} \quad (4.21)$$

we have that:

$$\hat{\theta}_{\mathrm{i}} \sim \mathrm{N}(\theta, (w_i^*)^{-1}) \tag{4.22}$$

If we treat the term $(w_i^*)^{-1}$ as it was the real variance of $\hat{\theta}_{\mathrm{i}}$ and not the estimated, then the test statistic we have is:

$$U^* = \frac{(\sum_{i=1}^{r} \theta_{\mathrm{i}} w_i^*)^2}{\sum_{i=1}^{r} w_i^*} \tag{4.23}$$

whereU^*follows a chi-square distribution with one degree of freedom, under the null hypothesis of no treatment difference (θ=0). If $(w_i^*)^{-1}$ was indeed the true variance of $\hat{\theta}_{\mathrm{i}}$, then the ML estimate of θ is:

$$\hat{\theta}^* = \frac{\sum_{i=1}^{r} \hat{\theta}_{\mathrm{i}} w_i^*}{\sum_{i=1}^{r} w_i^*} \tag{4.24}$$

Now $\hat{\theta}^*$ is asymptotically unbiased for θ, with variance to be approximately equal to

$$\frac{1}{\sum_{i=1}^{r} w_i^*} \tag{4.25}$$

Furthermore, the standard error is given by:

$$\mathrm{se}(\hat{\theta}^*) = \sqrt{\frac{1}{\sum_{\mathrm{i=1}}^{\mathrm{r}} w_i^*}} \tag{4.26}$$

and a 95%b CI for θ:

$$\hat{\theta}^* \pm 1.96 \sqrt{\frac{1}{\sum_{i=1}^{r} w_i^*}} \tag{4.27}$$

There is the possibility of having small $\hat{\tau}^2$. In that case the modified weights w_i^* will be close to the original ones. Moreover, the standard error and the CI derived from the random effects model will be similar to those derived from the fixed effect model. Plus, the overall estimate of treatment difference from both models will be similar. On the other hand, if $\hat{\tau}^2$ is large then we will have larger standard error and CI for the

random effects model. Furthermore, the random effects estimate of treatment difference will move closer towards the arithmetic mean of the individual study estimates. The difference between the estimate of random effects and the one of the fixed models will depend on the extent to which the studies with the largest original weights w_i are linked with the extreme estimates of treatment difference.

4.3.3: Estimation of τ^2 with the method of moments

The approach to the estimation of τ^2we are going to represent here is the one based on the method of moments. For this method a statistical software package is not needed.

When we have to deal with a random effects model, although the fixed model effects estimate of θ still has mean θ, its variance is given by:

$$var(\hat{\theta}) = \frac{\sum_{i=1}^{r} w_i^2 var(\hat{\theta}_i)}{(\sum_{i=1}^{r} w_i)^2} = \frac{\sum_{i=1}^{r} w_i^2 (w_i^{-1} + \tau^2)}{\sum_{i=1}^{r} w_i^2}$$
$$= \frac{1}{(\sum_{i=1}^{r} w_i)} + \frac{\tau^2 \sum_{i=1}^{r} w_i^2}{(\sum_{i=1}^{r} w_i)^2} \quad (4.28)$$

To test heterogeneity we use the statistic Q, given by:

$$Q = \sum_{i=1}^{r} w_i(\hat{\theta}_i - \hat{\theta})^2 = \sum_{i=1}^{r} w_i (\hat{\theta}_i - \theta)^2 - \left(\sum_{i=1}^{r} w_i\right)(\hat{\theta} - \theta)^2 \quad (4.29)$$

and the expected value of Q is:

$$E(Q) = \sum_{i=1}^{r} w_i var(\hat{\theta}_i) - \left(\sum_{i=1}^{r} w_i\right) var(\hat{\theta})$$
$$= \sum_{i=1}^{r} w_i (w_i^{-1} + \tau^2) - \left(\sum_{i=1}^{r} w_i\right)\left\{\frac{1}{\sum_{i=1}^{r} w_i} + \frac{\tau^2 \sum_{i=1}^{r} w_i^2}{(\sum_{i=1}^{r} w_i)^2}\right\}$$
$$= (r-1) + \tau^2\left(\sum_{i=1}^{r} w_i - \frac{\sum_{i=1}^{r} w_i^2}{\sum_{i=1}^{r} w_i}\right) \quad (4.30)$$

This motivates use of the method of moments estimate for τ^2, where

$$\hat{r}_2 = \frac{Q - (r-1)}{\sum_{i=1}^{r} w_i - \sum_{i=1}^{r} w_i^2 \Big/ \sum_{i=1}^{r} w_i} \quad (4.31)$$

as DerSimonian and Laird mentioned [18].

However, there is the possibility of a negative method of moments estimate. That is why, in practice, the estimate derives from the maximum of the values 0 and $\hat{\tau}^2$, i.e. when Q is smaller than the degrees of freedom the method of moments estimate will be 0.

4.3.4: A likelihood approach to the estimation of τ^2

As we have mentioned above, the random effects model has the below distributional assumption:

$$\hat{\theta}_i \sim N(\theta, \xi_i^2 + \tau^2) \quad (4.32)$$

In the following approach to the estimation of τ^2 we see that we treat to w_i^{-1}as it was known and equal to ξ_i^2. The contribution to the likelihood function derived from study I is:

$$L(\theta, \tau^2; \hat{\theta}_i) = \frac{1}{\sqrt{2\pi(w_i^{-1} + \tau^2)}} \exp\left\{\frac{-(\hat{\theta}_i - \theta)^2}{2(w_i^{-1} + \tau^2)}\right\} \quad (4.31)$$

When we have to deal with a meta-analysis that involves r independent studies, then the likelihood function is given by the product of the individual study likelihood functions. Plus, the log-likelihood function is given:

$$l(\theta, \tau^2; \hat{\theta}_i, i = 1, \dots, r) = constant - \frac{1}{2}\sum_{i=1}^{r} \log(w_i^{-1} + \tau^2) - \frac{1}{2}\sum_{i=1}^{r} \frac{(\hat{\theta}_i - \theta)^2}{(w_i^{-1} + \tau^2)} \quad (4.32)$$

We can find the ML estimates of τ^2 and θ through a repetitive scheme. On that scheme each of the iterations has two steps. On the first step, the variance parameter τ^2 is treated as fixed and we calculate the value of θ which maximizes the log-likelihood. On the second step, we treat θ as fixed and we calculate the value of τ^2 which maximizes the log-likelihood. Thus, the estimate of θ at the (t+1)th cycle is:

$$\theta^*_{i,t+1} = \frac{\sum_{i=1}^{r} \hat{\theta}_i w^*_{it}}{\sum_{i=1}^{r} w^*_{it}} \tag{4.33}$$

for t=0,1….., where $w^*_{it} = (w_i^{-1} + \hat{\tau}^2_{M,t})^{-1}$ and $\hat{\tau}^2_{M,t}$ is the ML estimate of τ^2 at the tth cycle of the iretation. In order to calculate the ML estimate at the (i+1)th cycle of the iretation, we use the Newton- Raphson procedure. Alternative, since it must satisfy the equation

$$\sum_{i=1}^{r} w^*_{i,t+1} = \sum_{i=w^*_{i,t+1}1}^{r} (w^*_{i,t+1})^2 \left(\hat{\theta}_i - \theta^*_{i,t+1}\right)^2 \tag{4.34}$$

We have the following approximate estimate

$$\hat{\tau}^2_{M,t+1} = \frac{\sum_{i=1}^{r} (w^*_{it})^2 \left\{\left(\hat{\theta}_i - \hat{\theta}^*_{i+1}\right)^2 - w_i^{-1}\right\}}{\sum_{i=1}^{r} {w^*_{it}}^2} \tag{4.35}$$

To start the iterative process the method of moments estimate of τ^2 could be used as the initial value $\hat{\tau}^2_{M,0}$.

Most of the times, the maximum likelihood estimate of τ^2 is underestimated because the method takes no account of the information used in estimating θ. However, the residual maximum likelihood (REML) actually takes account of that loss by modifying the likelihood equation to eliminate the parameter θ. The REML log-likelihood function is based on the residual terms instead of the observations $\hat{\theta}_i$, given by:

$$\ell_{\mathrm{R}}\{\tau^2;(\hat{\theta}_{\mathrm{i}}-\hat{\theta}^*_{\mathrm{i+1}})\mathrm{i}=1,\dots\mathrm{r}\}$$
$$=\text{constant}-\frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{r}}\log(\mathrm{w}_{\mathrm{i}}^{-1}+\tau^2)-\frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{r}}\frac{(\hat{\theta}_i-\hat{\theta}^*_{\mathrm{i+1}})^2}{(\mathrm{w}_{\mathrm{i}}^{-1}+\tau^2)}$$
$$-\frac{1}{2}\log\left\{\sum_{\mathrm{i}=1}^{\mathrm{r}}\frac{1}{(\mathrm{w}_{\mathrm{i}}^{-1}+\tau^2)}\right\} \tag{4.36}$$

We can calculate the REML estimate with a similar iterative scheme as the above, the difference is that now we have $\mathrm{w}^*_{\mathrm{it}}=(\mathrm{w}_{\mathrm{i}}^{-1}+\hat{\tau}^2_{\mathrm{R,t}})^{-1}$. We use the Newton-Raphson procedure in order to calculate the REML estimate of τ^2 at the (t+1)th cycle of the iteration. Otherwise, since it needs to satisfy the following equation:

$$\sum_{\mathrm{i}=1}^{\mathrm{r}}\mathrm{w}^*_{\mathrm{i,t+1}}=\sum_{\mathrm{i}=\mathrm{w}^*_{\mathrm{i,t+1}}1}^{\mathrm{r}}(\mathrm{w}^*_{\mathrm{i,t+1}})^2(\hat{\theta}_{\mathrm{i}}-\theta^*_{\mathrm{i,t+1}})^2+\frac{\sum_{\mathrm{i}=1}^{\mathrm{r}}(\mathrm{w}^*_{\mathrm{i,t+1}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{r}}\mathrm{w}^*_{\mathrm{i,t+1}}} \tag{4.37}$$

and approximate estimate is given by

$$\hat{\tau}^2_{\mathrm{R,t+1}}$$
$$=\frac{\sum_{i=1}^{r}(\mathrm{w}^*_{\mathrm{it}})^2\left\{\frac{r(\hat{\theta}_i-\hat{\theta}^*_{i+1})^2}{(r-1)-w_i^{-1}}\right\}}{\sum_{i=1}^{r}(\mathrm{w}^*_{\mathrm{it}})^2} \tag{4.38}$$

There are several ways to calculate the ML and REML estimates. Programs can be written or statistical packages can be used which fit multilevel models like MLn and SAS PROC MIXED.

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

CHAPTER 5

EXAMINING THE HETEROGENEITY

5.1: Introduction

Very often, as we have already mentioned in previous chapters, a meta-analysis is retrospective, i.e. the results are combined from studies which have not followed a common protocol. On the other hand, in a prospectively planned multicenter study, we often have a common protocol for the collection of key data. There is a continuum from the prospectively planned multicenter study to the retrospective conducted meta-analyses in terms of the validity of combining results. Generally, there will be less concern when we have to present combined results from a multicenter study compared with a meta-analysis in which different patient selection criteria, definition of the response measure or treatment regiments may have been used.

We do know that any mathematical model chosen and used for the conduction of a meta-analysis is only an approximation of the truth. It is crucial to choose models that not only fit to the data but are helpful for the interpretation of the results. Also, it is often to have a large number of analyses that has to be undertaken. Therefore, due to the fact that we must focus our analyses to a point, we must state the main analysis strategy a priori. Once we perform our main analysis, then secondary analyses may take place to help the implementation of the results or to address secondary issues.

A number of issues which needs to be mentioned in the specification of the main meta-analysis have to do with the heterogeneity in the treatment difference across trials. As we mentioned in previous chapters, we must decide whether the amount of heterogeneity is high or not, and if it is indeed high, the ways to deal with it.

In this chapter, we will present the use of a formal test for heterogeneity. In this test, factors are used in order to affect the choice between fixed and random effects model plus to decide whether or not it is appropriate to present an overall estimate of treatment difference.

In some cases, the treatment difference may be expected to vary from one level of a factor to another. As an example, we can think the large difference which may be expected in patients with a severe form of the disease than a mild form. Such factors are referred to as potential effect modifiers. However, it can be thought useful to state a priori a small number of factors as potential sources of heterogeneity. In that case, if the size of the treatment difference is indeed affected by the level if these factors, for example if there is a large number of effect in patients with severe form of the disease than a mild form, then the treatment difference can be presented for each group separately. The additional data, used to test the heterogeneity, might be at the study or the patient level. When the data from individual patient are available, the analysts have the possibility of adjusting for prognostic factors which are considered to be likely to affect the outcome data. For example, if we have to deal with trials of an antithypertensive agent blood pressure can be adjusted for the age of the patient, or in trials for Alzheimer's disease cognitive impairment may be adjusted for baseline disease severity [5].

5.2: The use of a formal test for heterogeneity

There is the possibility to perform a formal statistical test for heterogeneity across trials of the parameter which test the treatment difference. We have already mentioned statistical tests which can be performed to the case where study estimates of the treatment difference are to be combined in chapter 4.

Despite the fact that the result of a statistical test for heterogeneity does provide useful descriptive information about the variability among trials, any

decision based only on p-values is not recommended. A necessary distinction between a clinically important and a statistically significant difference must to be done. Usually, when we have to do with a single study, due to the design it follows, we do have the ability to detect a clinically important difference. However, there is an attempt to match the statistical along with the clinical significance. When we have a retrospective analysis, usually there is no control over the sample size, so it is better to consider the variation in the size of the effect which would be considered clinically important. When we have a large amount of data set, then an amount of heterogeneity may be statistically important compared to a small data set, where an amount of heterogeneity may not be statistically important. In the random effects model we presented earlier, the study treatment difference parameters are assumed to be independent, normally distributed observations from $N(\theta, \tau^2)$. Therefore, the coefficient of variation τ/θ can be though as a useful additional measure and it can be estimated by substituting estimates of θ and τ into the numerator and denominator, respectively.

In 1998, Hardy and Thompson tried to investigate the efficiency and the power of the test of heterogeneity based on the statistical test Q under different scenarios[22]. Those scenarios included varying the size of the heterogeneity parameter, the number of trials taking part in the meta-analysis (r), and the weight w_i of the ith study, for i-1,…r. They came to the conclusion that the power can be low, especially when a trial has been weighted much larger than the others. Furthermore, they added that the test of heterogeneity for a fixed effects model is of limited use, especially when the total information is low or when we have a large variability among the weights of the trials.

5.3: The choice between a fixed effects and a random effects model

When an analyst must decide whether to use fixed or random effects model, he/she must not be based solely on the results of the test of heterogeneity.

Additional criteria, such as the number of trials or the distribution of the study estimates of treatment difference, must to be considered. As we have already mentioned in section 6.2 when we have to deal with a small number of studies, then the estimate of the heterogeneity parameter may be unreliable. However, if the results from the trial seem to be reasonable enough, then the fixed effects may be more appropriate to be used. On the other hand, in the case where the results seem to be unreasonable, then there is no need to calculate an overall estimate, but to investigate more the cause of this inconsistency. When the meta-analysis is based on large number of trials, then the random effects analysis is more appropriate and we will discuss it further in following paragraphs. However, in that case, if the distribution of the estimates seems to be far away from the assumption of normality, then further investigation must to be done.

The overall estimate of the fixed effects model provides a summary of the results, which are obtained from the particular sample of patients contributing data. When we extrapolate the results from the fixed effects model to the total population, we make the assumption that the characteristics of the patients who contribute to the meta-analysis are the same to those in the total patient population. On the other hand, the random effects model produces results which can be thought to be more generalized. However, the basic assumption of the random effects model is that the results from the studies in the meta-analysis are representative of the results which would be obtained from the total population of the treatment centers plus that the centers are usually chosen at random. Despite that alternatives can be considered, the normal distribution is chosen for modeling the heterogeneity in the treatment difference parameter across trials because of its robustness and the computational convenience. One advantage of the random effects model which must to be pointed out is that this model allows the between-study variability in the treatment difference estimates to play part in the overall estimation by influencing the results and the precision of them. The above, allows the substantial variability, if it exists, to be reflected in a wide confidence interval. The disadvantage of this approach is that the demonstration of a significant treatment benefit, compared to the fixed effects

https://doi.org/10.26219/heal.aueb.6071

model, requires a large number of participant patients. As a result, definitive evidence of treatment efficacy from a random effects model will be more convincing.

The best approach in most cases is to consider the results from both fixed and random effects model. If there is no heterogeneity, then the two models will provide the same results, as the estimate of τ^2 will be equal to 0. However, if there is significant difference between the two models then this will lead to conclusion that more investigation must to be done.

If the study estimates have substantially differences then is quite inappropriate to represent an overall estimate. Let's see an hypothetical example concerning three studies comparing a new drug against placebo (Table 5.1).

Study	$\hat{\theta}_i$	95% CI	w_i	$\hat{\theta}_i w_i$	$\hat{\theta}_i^2$
1	0.6	(0.1,1.0)	22	13.2	7.9
2	3,0	(2.5,3.5)	15	45.0	135.0
3	0.5	(0.1,0.9)	30	15.0	7.5
Total			67	73.2	150.4

$U=(73.2)^2/67=80.0$;(1 df) $p<0.001$

$\hat{\theta} = 73,2/67=1.09$;se($\theta\hat{}$) $=1/\sqrt{67}=0.12$; 95% CI= (0.85, 1.33)

$Q=150.4-80.0=70.4$; (2df) $p<0.001$

$\hat{\tau}^2 = (70.4-2)/(67-1609/67)=1.59$

$U^* (2.487)^2/1.831=3.37$; (1df) $p=0.07$)

$\hat{\theta}^* = 2.487-1.831=1.36$; se($\theta\hat{}^*$)$=1/\sqrt{1.831}=0.74$; 95% CI=(-0.09, 2.81)

Table 5.1: Meta-analysis of three studies comparing a new drug with placebo. The measure of treatment difference is denoted by θ, which is positive when the new drug is beneficial.

The results, if we examine them individually, show that all three new drugs have, on their favor, statistically significant against the placebo. Both studies 1 and 3 have a similar size effect, compared to the 2nd one where the size effect is bigger. Consider a meta-analysis which combines the study estimates from those three studies using the methods we discussed in an earlier chapter.

The 95% CI[4] which is based on the fixed effects methods model lies between the two extremes but is not consistent with either, so it does not seem an appropriate summary of the results. Moreover, the test for heterogeneity when we use the Q statistic is highly significant. On the contrary, when we use the random effects model we observe that the treatment difference is not significantly different from 0 at the 5% level, along with the fact that the CI is much wider, including small negative values[5]. That is also not seems appropriate to be used. However, it is more than welcome to investigate more the studies, especially the fact that the effect of the 2nd study is different from the other two.

From the above table, we can see that all three studies show a statistically significant benefit for the new drug against the placebo. However, it should be more concern about the variation in the size effect. If we use the random effects model, which calculates the probability that the new drug is worse than the placebo at a fourth center, i.e. $1-\Phi(1.36/\sqrt{1.59})$ where Φ is the standard normal distribution function, we see that the resulting value of 0.14 is not reassuringly small.

In that point, we must underline the distinction between the quantitative and qualitative interaction. The quantitative interaction is the term which we apply to heterogeneity between studies when the effects are either all negative or all positive. The qualitative interaction shows that the drug is either beneficial in some cases or harmful in some others. Between those two, more concern is given if qualitative interaction happens. The above example shows a kind of qualitative interaction. However, if we subtract 2 from all of the study estimates and CIs, we obtain a qualitative interaction. In this case, studies 1 and 3 would have individually concluded a significant effect in favor of placebo, when study 2 still shows a significant effect in favor of the new drug.

[4]The CI for the fixed effects model is (0.85, 1.33).
[5]The CI for the random effects model is (-0.09, 2.81).

5.4: Meta-regression using study estimates of treatment difference

We have the chance to explore the dependence of the treatment difference on one or more characteristics using a meta-regression. Meta-regression is an analysis in which the trials estimates of treatment difference are the observations and trial level covariates, which they have all a value defined for each trial, are the exploratory variables. However, especially when there is a large number of studies, it would be more preferable to examine one factor at the time.

In order to incorporate a trial level covariate within the fixed effects model we use the following formula:

$$\hat{\theta} = \beta_1 + \eta_i + \varepsilon_i \qquad (5.1)$$

where $\eta_i = \beta_2 x_{2_i}$ in the case of a continuous exploratory variable x_{2_i} and $\eta_i = \beta_2 x_{2_i} + \beta_3 x_{3_i} + \cdots + \beta_q x_{q_i}$ in the case we have a factor with q levels. Here, $x_{2_i}, \ldots, x_{q_i}$ are a set of q-1 indicator variables, which can take only two values, 0 and 1. The ε_i are the error terms, which are normally distributed random variables with expected values 0 and variance ξ_i^2. In the case we had no exploratory variables, the β_1 would be equal to θ.

In order to calculate the maximum likelihood estimates of the β_s, we can perform a weighted least- squares regression in $\hat{\theta}_i$ on the exploratory variables, with weights w_i[6]. We can calculate the maximum likelihood estimates using any statistic package. However, despite the fact that most of the packages present in their output the correct estimates of the regression coefficients, the standard errors and test statistics are incorrect for the meta-analysis model. This happens due to the fact that most of the packages assume that $\xi_i^2 = {\sigma^2}/{w_i}$, where σ^2 supposed to be estimated from the data, instead of $\xi_i^2 = {1}/{w_i}$. If we want the

[6] The w_i is the estimated inverse variance of $\hat{\theta}_i$

correct standard error for $\hat{\beta}_j$, then we must divide the $\hat{\beta}_j$ given from the package with the square root of the residual (error) mean square. Another way to obtain the correct standard errors is the square roots of the diagonal elements of the matrix $(X'WX)^{-1}$, where X is the rxq matrix of exploratory variables associated with the βs and W is the rxr diagonal matrix with ith element w_i.

If the model includes only the term β_i, then the estimate of β_i is the overall fixed effect estimate $\hat{\theta}$. Additionally, U and Q statistics appear as the model sum of squares and residual (error) sum of squares respectively in the analysis of variance table. When we add an exploratory variable in the model, then the Q statistic is divided in two terms: the Q_B and Q_W. The first is the variation explained by the covariate and the second one is the remaining unexplained variation, which appears as the residual sum of squares.

If the exploratory variable is as we said earlier with q levels, then in order to test the heterogeneity on the treatment difference parameter between the studies with the same factor level, we compare Q_W with the chi-squared distribution with r-q degrees of freedom and Q_B with the chi-squared distribution with q-1 degrees of freedom. We have:

$$Q_B = \sum_{k=1}^{q} \left\{ \frac{\left(\sum_{i=1}^{n_k} w_{ki}\hat{\theta}_{ki}\right)^2}{\sum_{i=1}^{n_k} w_{ki}} \right\} - \frac{\left(\sum_{k=1}^{q}\sum_{i=1}^{n_k} w_{ki}\hat{\theta}_{ki}\right)^2}{\sum_{k=1}^{q}\sum_{i=1}^{n_k} w_{ki}} \tag{5.2}$$

where $\hat{\theta}_{ki}$ is the estimate of treatment difference from the ith study at the kth level of the factor and w_{ki} its weight, for i=1,…,n_k and k=1,…,q.

If the exploratory variable is a continuous one, then Q_B is compared with the chi-squared distribution with one degree of freedom. Then, we have:

$$Q_B = \frac{\left\{\sum_{i=1}^{r} w_i(x_{2i} - \bar{x}_2)\hat{\theta}_i\right\}^2}{\sum_{i=1}^{r} w_i(x_{2i} - \bar{x}_2)^2} \tag{5.3}$$

where

https://doi.org/10.26219/heal.aueb.6071

$$\bar{x}_2 = \frac{\sum_{i=1}^{r} V_i x_{2i}}{\sum_{i=1}^{r} V_i}$$

5.5: A strategy for dealing with heterogeneity

Heterogeneity is a factor that must be always be evaluated in every meta-analysis, the investigation of which can be separated in two parts. The first part is specified a priori in the protocol and the second one is an additional approach which may or may not be needed.

Questions that might be addressed in the protocol include:

a) the smallest treatment different which can be considered to be clinically important
b) the statistic which we will use to test the heterogeneity
c) study-level covariates to be included in the meta-analysis
d) patient-level covariates which will be adjusted for imbalance in the distribution of specific prognostic factors and baseline characteristics across treatment groups
e) patient- level covariates to be evaluated as potential effect modifiers.

In the case where the heterogeneity is considered to be clinically important but is difficult to be explained by the sources of potential heterogeneity specified above, then extra expiratory analyses, in which we must include more covariates, may be needed. Another step that must be taken is to check whether the covariates that have been used so far are appropriate. For example, in the case where the data is binary, what is the best to use: log-odds ratio or probability difference? Analysis of other related variables will show whether or not the heterogeneity is restricted only to the primary response variable. If eventually there is no reason to explain the heterogeneity, then there is a good step to use a random effects model, where there is the possibility for the treatment difference to vary from study to study.

https://doi.org/10.26219/heal.aueb.6071

5.6: Examples

5.6.1: Global impression of change in Alzheimer's disease

Let's see the data from the tacrine studies we described in the section 3.4. The following tables provide us with the necessary data for our calculations:

Study	$\hat{\theta}_i$	se($\hat{\theta}_i$)	95% CI
1	0.284	0.261	(-0.228, 0.797)
2	0.224	0.242	(-0.251, 0.699)
3	0.36	0.332	(-0.290, 1.011)
4	0.785	0.174	(-0.444, 1.126)
5	0.492	0.421	(-0.334, 1.318)

Table 5. 2: Study estimates of the log-odds ratio from the proportional odds model for the tacrine studies, based on the maximum likelihood approach.

Study	Treatment	Category					$\hat{\theta}_i$	w_i	$\hat{\theta}_i w_i$	$\hat{\theta}_i^2 wi$
		C1	C2	C3	C4	C5				
1	Tacrine	4	23	45	22	2	0.284	14.63	4.16	1.18
	Placebo	2	22	54	29	3				
2	Tacrine	14	119	180	54	6	0.224	17.02	3.81	0.85
	Placebo	1	22	35	11	3				
3	Tacrine	13	20	24	10	1	0.6	9.08	3.27	1.18
	Placebo	7	16	17	10	3				
4	Tacrine	21	106	175	62	17	0.785	33.03	24.92	20.34
	Placebo	8	24	73	52	13				
5	Tacrine	3	14	19	3	0	0.492	5.63	2.77	1.36
	Placebo	2	13	18	7	1				
Total								79.41	39.94	24.92

U=(39.94)2/79.41=20.09;(1 df) p<0.001
Q=24.92-20.09=4.83; (4df) p=0.30
$\hat{\theta}$ = 39.94/79.41=0.503;se(θˆ) =1/√79.41=0.112;
95% CI= (0.503±1.96√79.41)=(0.283 , 0.723)

Table 5. 3: Fixed effects meta-analysis of the log-odds ratio from the proportional odds model for the tacrine studies based on the results from table 5.2

The test for heterogeneity across the studies described on table 5.3 was not statistically significant (p=0.30). However, it was for our interest to investigate more the effect of the dose of tacrine on the treatment different. However, the relationship between the log-odds ratio and dose was difficult to be assessed, due to the fact that in most studies the dose each patient was receiving was titrated or selected to be the patient's best dose. The final dose which was

actually received by the patient in a trial was thought to be a measure of the intended level or dosing for the trial. These doses were 62, 39, 66, 135 and 65mg/day for studies 1 to 5 respectively. We observed that as the dose was increasing, the treatment effect was increasing too. We consider that the dose is a continuous variable in the meta- regression. The following table shows the residual sum of squares from fitting the null model[7] which is equal to the value of Q given in the table 5.3 (the value 4.83).

Model	Residual sum of squares	Degrees of freedom
Intercept only	4.83	4
Dose	0.15	3

Table 5. 4: Global impression of change in Alzheimer's disease: meta-regression of the log-odds from the proportional odds model on the dose of tacrine

When we include in our model the covariate for dose we can see a substantially reduce of the residual sum of squares from 4.83 to 0.15, by 4.68. Now, if we compare the 4.68 with chi-squared distribution with one degree of freedom, we have p-value equals to 0.03. This value indicates that the dose increases significantly when we use the log-odds ratio.

Parameter	Estimate	Standard error	95 % CI
b_1	-0.023	0.268	(-0.549, 0.52)
b_2	0.00597	0.00276	(0.00056, 0.01139)
Dose 65mg/day (b_1+65b_2)	0.365	0.129	(0.112, 0.618)
Dose 135mg/day (b_1+135b_2)	0.783	0.171	(0.447, 1.119)

Table 5. 5: Parameter estimates from the meta-regression of the log-odds ratio from the proportional odds model in the dose of tacrine.

[7] That is the model with the intercept term only.

The estimates of the log-odds from the meta-regression for the dose 65 and 135 show us that there is an agreement with the individual study estimates. The residual sum of squares after the fitting dose is very small. If we compare the chi-squared distribution with three degrees of freedom, the value of 0.15 is not statistically significant p=0.98. For the ML and the REML estimates of the residual variance component after fitting the dose we have that they were equal to 0.

5.6.2: Recovery time after anaesthesia

We illustrated the anaesthetic study in section 3.3.3. The following table shows that there is heterogeneity between the individual centre estimates.

Center	Treatment A		Treatment B						
	n	Mean	n	Mean	Standard Deviation	$\hat{\theta}_i$	w_i	$\hat{\theta}_i w_i$	$\hat{\theta}_i^2 wi$
1	4	1.141	5	0.277	0.62	0,864	4,4	3,8	3,28
2	10	2.165	10	1.519	0.913	0,646	9,89	6,39	4,12
3	17	1.79	17	1.518	0.849	0,272	16,81	4,58	1,25
4	8	2.105	9	1.189	1.061	0,916	8,380	7,680	7,030
5	7	1.324	10	0.456	0.619	0,867	8,15	7,07	6,13
6	11	2.369	10	1.55	0.558	0,819	10,36	8,49	6,95
7	10	1.074	12	0.265	0.502	0,809	10,79	8,73	7,06
8	5	2.583	4	1.37	0.934	1,212	4,4	5,33	6,46
9	14	1.844	19	2.118	0.749	-0,273	15,95	-4,35	1,19
Total							89,12	47,69	43,48

U=(47.69)2/89.12=25.52;(1 df) p<0.001
Q=43.48-25.52=17.95; (8df) p=0.02
$\hat{\theta}$ = 47.69/89.12=0.535;se(θˆ) =1/√89.12=0.106;
95% CI= (0.503±1.96√89.12)=(0.328, 0.743)

Table 5. 6: Fixed effects meta-analysis of the absolute mean difference (treatment A- treatment B) in log- recovery time, assuming a common variance across all centers.

We observe that in the center 9 there is a negative mean difference, which indicates that anaesthetic A is able to reduce recovery time compared to anaesthetic B, whereas in the other eight centers there is the reverse case. Moreover, the test of heterogeneity was statistically significant with p- value equals to 0.02. Moreover, further study of the results showed that the centers

were able to choose the premedication drug administered, provided that all the patients received the same drug at the center. Centers 1 to 8 used the same drug, in contrast with the center 9. We can create a study- level covariate from the use of the premedication drug. Let this covariate take the value 0 for premedication 1, which was used in centers 1 to 8, and 1 for premedication 2 which was used in center 9.

Model	Residual sum of squares	Degrees of freedom
Intercept only	17.95	8
Dose	5.27	7

Table 5. 7: Recovery time after anaesthesia: meta-regression of the absolute mean difference on premedication

From the above table, we can see that the residual sum of squares from fitting the model only with the intercept equals to the value of Q, as we can see from the table 5.6. If we include the covariate in that model, we observe that the residual sum of squares reduces from 17.95 to 5.27, losing 12.68 units. That reduction of 12.68 compared with the chi-squared distribution with one degree of freedom gives p-value less than 0.001. Furthermore, the data from the below table, shows that anaesthetic B reduces recovery time relative to anaesthetic A, when we use the premedication 1 but there is some evidence that anaesthetic B is better than A, despite the fact that is not statistically significant. Moreover, the same results with the meta-regression would be obtained using the fixed effects model for centers 1 to 8, conditionally that the same σ^2 is used.

The residual sum of squares after the premedication can be compared with the chi-squares distribution with 7 degrees of freedom. The value 5.27 which is statistically significant since the p-value equals to 0.63 indicates that there is no strong evidence of heterogeneity among the first eight studies. The ML and REML estimates of the residual variance component after fitting premedication were 0. After the meta-regression we can conclude that the choice of anaesthetic agent depend on the premedication which is going to be used. For premedication 1, anaesthetic B has fast recovery time, but for the premedication

https://doi.org/10.26219/heal.aueb.6071

2 there is no clear result but there some indication that the anaesthetic A might be better.

https://doi.org/10.26219/heal.aueb.6071

CHAPTER 6

DEALING WITH NON-STANDARD DATA SETS

6.1: Introduction

In some meta-analyses, the characteristics of the available data make it difficult, and in some cases even impossible, to implement the methods we have already analyzed in earlier chapters. In this chapter, we are going to discuss the solutions in some very common problems in meta-analyses.

Even if all studies which are included in a current retrospective meta-analysis have the same outcome measure, the way in which the results are presented in a published report or paper may vary among the studies. We will see ways of combining trials which report with different summary statistics. Moreover, there are meta-analyses which are planned to be performed using individual patient data, but it may only be possible to obtain summary information from some studies. In that case, the analyst must combine estimates of treatment difference based on summary statistics with those based on individual patient data. Furthermore, there are methods of combining p-values when we have the possibility to calculate estimates of treatment difference from individual studies.

The above are going to be discussed in this chapter.

6.2: Combining trials which report different summary statistics

In some cases, the way summary statistics for a particular outcome measure are reported may vary from one study to another, due to different undertaken methods of analysis. However, this is able to create a problem if the meta-analysis is based on summary information from published papers, the

extension of which depends on the type of outcome measure which is combined. For example, usually there is not a problem for binary data, since usually there is sufficient information of the categories for each treatment group. However, if we have to deal with ordinal data with more than two categories, where the number of the patient for each category is usually not provided, then it could be a problem.

6.2.1: Continuous outcomes

Very often, an outcome measured on a continuous quantitative scale is treated as it is derived from a normal distribution. The summary statistics presented in the majority of the published papers have to do with the number of patients, the sample mean and the standard deviation for each group participating on the specific study. However, the summary information deriving from a continuous data often is reported as if it is related to a binary outcome. For instance, a patient can be reported as a responder if a particular value on the continuous scale is exceeded and a non-responder otherwise. If an analyst wants to combine summaries of binary outcomes with those of continuous outcomes, he can chose as a common measure of treatment difference the log-odds ratio. More details of this methodology can be found in the paper of Whitehead(1999) [19]. It is illustrated by a series of perinatal trials investigating the effect of prophylactic use of oxytocics on postpartum blood loss during labor. Moreover, one of the meta- analysis presented in the paper, which concerns the combination of eight trials reporting binary outcomes along with three reporting continuous outcomes, is considered here.

In the above example, the binary outcome in the perinatal trial has to do with whether a woman had to do a postpartum haemorrhage, which is usually defined by a blood loss of 500ml or more in the first 24 hours after delivering the baby. The continuous outcome we mentioned had to do with the actual amount of

blood lost. The log-odds ratio of a postpartum haemorrhage on the oxytocic treatment relative to the control treatment is:

$$\theta = log\left\{\frac{p_T(1-p_C)}{p_C(1-p_T)}\right\} \tag{6.1}$$

where p_T and p_C are the probabilities of a haemorrhage in the oxytocic and control groups respectively. When we have to deal with summary information with binary outcome, the log-odds ratio and its variance are estimated from (3.2) and (3.3).

On the other hand, for the continuous outcome, the reported summary statistics are the number of patients, mean and standard deviation in each treatment group. Let Y_T and Y_C be the continuous outcome variables in one trial for the oxytocic and control treatment, respectively. We assume that the individual patient observations are normally distributed, with $y_{T_j} \sim N(\mu_T, \sigma^2)$, j=1,….,$n_T$ and $y_{C_j} \sim N(\mu_C, \sigma^2)$, j=1,…,$n_C$. Furthermore, we denote with A the cut-point value so that $p_T = P(Y_T > A)$and $p_C = P(Y_C > A)$. The ML estimate for $p_T$is $1 - \Phi(A_T)$, where $A_T = \frac{(A-\bar{y}_T)}{\hat{\sigma}_M}$ and Φ is the standard normal distribution function. The statistics $\bar{y}_T$ and $\hat{\sigma}_M$ are the ML estimates of μ_T and σ^2 respectively. Similarly, we define the estimate of p_C. We have the ML estimate:

$$\hat{\theta} = \log\left[\frac{\Phi(A_C)\{1-\Phi(A_T)\}}{\Phi(A_T)\{1-\Phi(A_C)\}}\right] \tag{6.3}$$

The variance of $\hat{\theta}$ is obtained by the delta method and is given by:

$$\mathrm{var}(\hat{\theta}) = \frac{\{\varphi(A_T)\}^2\left(\frac{1}{n_T} + \frac{A_T^2}{2n}\right)}{[\Phi(A_T)\{1-\Phi(A_T)\}]^2} + \frac{\{\varphi(A_C)\}^2\left(\frac{1}{n_C} + \frac{A_C^2}{2n}\right)}{[\Phi(A_C)\{1-\Phi(A_C)\}]^2}$$

$$- \frac{A_T A_C \varphi(A_T)\varphi(A_C)}{n\Phi(A_T)\Phi(1-A_T)\Phi(A_C)\Phi(1-A_C)} \tag{6.4}$$

where$n = n_T + n_C$and φ is the standard normal density function.

https://doi.org/10.26219/heal.aueb.6071

The following table 6.1 shows the summary statistics from the 11 trials.

a) Binary outcomes

Trial	Oxytocic		Control	
	Haemorrhage	Total	Haemorrhage	Total
1	45	490	80	510
2	1	150	5	50
3	14	591	4	177
5	24	963	25	470
9	34	346	42	278
11	50	846	152	849
12	0	10	1	15
13	14	705	60	724

b) Continuous outcomes

Trial	Oxytocic			Control		
	Number	Mean	Standard deviation (s_T);df	Number	Mean	Standard deviation (s_T);df
6	41	150.49	86.31;35	10	305	59.86;9
8	97	188.35	84.18;95	43	216.65	119.35;42
10	319	125.14	97.68;317	122	233.2	107.40;121

Table 6.1: Prophylactic use of oxytocics on postpartum haemorrhage: summary statistics for each trial.

In all trials reporting binary outcomes, except trial 1, a postpartum haemorrhage was defined as a blood loss of 500 ml or more. In trial 1 we use a cut point value of 20 oz, which converts to 568 ml. On the other hand, the one of the continuous outcomes, the standard deviation presented in the published papers was assumed to be the usual unbiased estimate than the ML estimate. For a trial consisting of more than one oxytocic group, the results were pooled to provide one oxytocic group.In the case of the continuous data, this means assuming a common mean and variance for each oxytocic treatment. The degrees of freedom (fd) show the denominator for the calculation of the pooled variance.

The following table 6.2 shows the estimates of the percentage of women experiencing a haemorrhage and the log-odds ratio from each trial.

https://doi.org/10.26219/heal.aueb.6071

Trial		Estimated % of women experiencing a haemorrhage		Log-odds ratio	Std. error	95% CI
		Oxytocic	Control			
	1	9.2	1537	-0.61	0.2	(-1.00,-0.22)
	2	0.7	10	-2.81	1.11	(-4.98,0.63)
	3	2.4	2.3	0.05	0.57	(-1.08,1.17)
	5	2.5	5.3	-0.79	0.29	(-1.36,-0.22)
	6	<0.01	0.5	-7.84	1.88	(-11.52,-4.15)
	8	0.05	0.1	-0.91	0.63	(-2.13,0.32)
	9	9.8	15.1	-0.49	0.25	(-0.97,-0.01)
	10	<0.01	0.4	-3.75	0.42	(-4.58,-2.93)
	11	5.9	17.9	-1.24	0.17	(-1.58,-0.91)
	12	0	6.7	-0.78	1.68	(-4.07,2.52)
	13	2	8.3	-1.5	0.3	(-2.09,-0.90)
Fixed effects estimate				-1.08	0.09	(-1.26,-0.89)
Random effects estimate				-1.38	0.31	(-1.99,-0.77)
Test for heterogeneity (x^2)				74.79;(10df)	<0.001	

Table 6.2: Meta-analysis of the log-odds ratio of a haemorrhage on oxytocics relative to control, using the methods of Chapter 4 with the method of moments estimate of τ^2

We must point out that in trial 12 the ML estimate could not be calculated due to the fact that there were not haemorrhages in the oxytocic group. In order to include this specific trial in our analysis, we obtained to have an ML estimate by adding 0.5 to all cells in the 2x2 table. Using the methods of chapter 4 and the method of moment of τ^2 we performed both fixed and random effects meta-analyses. Comparing the estimates of the percentages of women experiencing a haemorrhage of the summary statistics, we can see that in the three continuous trials are much smaller than those on the other trials. This may be caused because few or no women experienced a blood loss less than 500 ml on these trials. In order to present a measure of treatment difference, the authors may have resorted to reporting the continuous outcome. Moreover, two out of three log-odds ratio have larger magnitude compared to the ones based on the binary outcomes. However, despite the doubt over the magnitude of the treatment difference, after analyzing the meta-analysis the conclusion that may be drawn from the meta-analysis is that the routine use of oxytocic drugs is beneficial in reducing the risk of major bleeding in the third stage of labor.

6.2.2. Ordinal data

Generally we can report an ordinal data in many different ways. For instance, we can report an ordinal data as if it were a binary outcome. For example, when we have a "success" we may constitute a response in the best category or in some cases in one of the best categories. The same ordinal outcome may be recorded in each study, however there is the possibility that the definition of "success" may vary from one study to another. If, in each study, we can report the number of patients in the "success" and "failure" categories for each treatment group, then the meta-analysis can be performed on the log-odds ratio. We must point out in this point that a meta-analysis based only on binary data may be making an implicit assumption of proportional odds in case that the definition of "success" is not the same among all studies. Furthermore, if some studies are able to provide the number of responses in more than two categories for each treatment group, then the analysts can calculate the log-odds ratio using a proportional odds model.

There are times where an ordinal data can be treated as if it were continuous one arising from a normal distribution. In that case, the summary statistics are the same with the ones described above, in 6.2.1. However, this is appropriate and can be used only when the difference between two consecutive scores is of equal clinical importance throughout the scale along with an approximately normally distributed data. If this is the case, then we use again the approach described in 6.2.1. It is a frequent case in meta-analyses to combine studies some of which report binary summary statistics and some continuous ones. The meta-analysis could proceed as for the oxytocic example. However, it is rather logical to consider using a logistic distribution rather than a normal distribution to model the continuous data. Despite the fact that those two distributions are similar, the logistic regression has the proportional odds property. The proportional odds property means that the log-odds ratio remains constant across all cut-points. However, we must point out that any extraction of relevant data pertaining to the logistic distribution is likely to be problematic.

6.3: Combining summary statistics and individual patient data

A meta-analysis based on individual patient data can be more reliable compared to the one using summary statistics. However, the basic problem arises when the individual data is not available for one or more trials. Those trials may be incorporated into the meta-analysis provided that sufficient summary information is presented in the trial report. In some cases, it would be desirable to perform the meta-analysis as a sensitivity analysis, although in some cases this might be the primary target either way.

In order to use the methodology we described in chapter 4, it is crucial to calculate an estimate of the chosen parameter measuring treatment difference, along with an estimate of its variance. For some data, as for example a binary data, there may be sufficient summary information to use all the above methodologies. As for normally distributed data, Goldstein et al (2000) presented a model for combining individual patient data with study level data.

6.4: Combining p-values

When we have to do with a typical meta-analysis, we know that we must have a calculation of study estimates of treatment difference along with an overall estimate. However, in some cases we may not have sufficient data to proceed in such calculations, especially when the only source of information we have is by published papers. A way to overpass this problem is to use methods, developed during the 1930s, for the combination of p-values, provided that these have been reported. This approach may also be taken if different outcomes measures have been reported from one study to the next and the assumptions required for either of the two approaches discussed earlier are not met.

Methods which derived on summarizing the p- values are based on one-sided p-values. To understand this approach we can consider the situation in which there is a common parameter, let's say θ, measuring the treatment

difference in all studies. We assume that θ equals to 0 when the two treatments are equivalent and takes positive values if the new treatment is a better one compared to the control one. Our interest is to test the null hypothesis that θ equals to 0 against the one-sided alternative that θ is greater than 0. The one-sided p-value is the probability to get a test statistic at least as extreme as the calculated in favor of this one-sided alternative given that the null hypothesis is true. Furthermore, let p_{1i} be the one sided p-value for study i. Often, the presented p-value in a trial report or publication is not usually p_{1i} but is more common to report p_{2i}, which is the p-value referring to the two-sided alternative that θ is not equal to 0. The value of p_{1i} can be calculated from p_{2i}. However, we must be very careful with this calculation. First of all, we must check whether the estimate of θ is positive or negative. In the case we have positive estimate then $p_{1i} = \frac{p_{2i}}{2}$. Otherwise, in the case of a negative estimate we have$p_{1i} = 1 - \frac{p_{2i}}{2}$. When we want to combine the p-values from different outcome measures, it is important to check that p_{1i} relates to the one-sided alternative that the new treatment is better than the controlled one.

Another assumption of the method for combining p-values is that p-values is a continuous variable, i.e. it can take all values between 0 and 1. In 1932, Fisher created a chi-squared statistic based on the p_{1i} in order to test the global null hypothesis that the two treatments are equivalent against the one-sided alternative that in at least one study the new treatment is better than control[20]. Under the null hypothesis p_{1i}in uniformly distributed between 0 and 1. Thus, the statistic $T_i = -2\log(p_{1i})$ has a chi-squared distribution with two degrees of freedom. We have:

$$P(T_i > t) = P(-2\log(p_{1i}) > t) = P\left(p_{1i} < \exp\left(-{}^{t}/_{2}\right)\right)$$
$$= \exp\left(-{}^{t}/_{2}\right) \quad (6.5)$$

Due to the fact that the r studies are independent, if the null hypothesis is true for each study, then we have that:

$$P = \sum_{i=1}^{r} T_i \tag{6.6}$$

follows a chi-squared distribution with 2r degrees of freedom.

In that point, we want to test the null hypothesis that the two treatments are equivalent against the one-sided alternative that in one at least study the old treatment is better than the new one, the test statistic

$$P^{-} = -2 \sum_{i=1}^{r} \log (1 - p_{1i)} \tag{6.7}$$

is compared with the chi-squared distribution with 2r degrees of freedom.

For the above method we will illustrate an example taken from Canner (1987) [26]. In this example, we have three major clinical trials where we compare the aspirin with placebo in coronary heart disease. The meta-analysis presented is based on the unconditional maximum likelihood estimation of the log-odds ratio for mortality on aspirin compared to placebo (formulae (3.2) and (3.3)).

We have the following table:

https://doi.org/10.26219/heal.aueb.6071

Study	Aspirin		Placebo		$\hat{\theta}_i$	w_i	$\hat{\theta}_i w$	$\hat{\theta}_i^2 w_i$
	Number of deaths	Total number patients	Number of deaths	Total number patients				
1	49	615	67	624	-0.329	25.7	-8.46	2.78
2	44	758	64	771	-0.385	24.3	-9.34	3.59
3	27	317	32	309	-0.216	13.3	-2.86	0.62
4	102	832	126	850	-0.220	18.8	-10.71	2.35
5	85	810	52	406	-0.225	28.4	-6.41	1.44
Total (1-5)						140.5	-37.78	10.78
6	246	2267	219	2257	0,125	104	12.96	1.62
Total (1-6)						244.5	-24.82	12.4

Studies 1-5
U=(-37.78)2/140.5=10.16;(1 df) p=0.001
Q=10.78-10.16=0.63; (4df) p=0.96
$\hat{\theta}$ = -37.78/140.5=-0.269;se(θˆ) =1/√140.5=0.084
95% CI= (-0.269±1.96/√140.5)=(-0.434 , -0.104)
Studies 1-6
U=(-24.82)2/244.5=2.52;(1 df) p=0.11
Q=12.40-2.52=9.88; (5df) p=0.08
$\hat{\theta}$ = -24.82/244.5=-0.102;se(θˆ) =1/√244.5=0.064
95% CI= (-0.102±1.96/√244.5)=(-0.227 , -0.024)

Table 6. 3: Aspirin in coronary heart disease: log-odds ratio of mortality on aspirin relative to placebo

The first five trials are very close to each other. We do not have heterogeneity (the test of heterogeneity had p-value=0.96) and the overall test of a treatment difference is significant with p-value=0.001. However, when we insert to the meta-analysis the sixth study the results change immediately showing that there is a higher level of mortality on aspirin than in the placebo. Due to the fact that this study is larger than the others, we observe a big reduction in the positive effect, in the point where the inclusion is not statistically significant (p-value= 0.11). Canner observing this huge difference when he inserted study six in the meta-analysis, he presented his findings.

The first which could cause this heterogeneity is the different design and operation features of the six trials. The following table shows some basic features of every trial:

	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
Time period	1971-73	1972-75	1970-77	1975-79	1975-79	1975-79
Number of patients	1126	1529	626	1682	1216	4524
Mean age	55,0	56,5	58,9	56,2	56,3	54,8
Gender	M	M	M,F	M,F	M,F	M,F
Total daily dose (mg)	300	972	1500	900	972	1000
Dosage schedule	o.d.	t.i.d.	t.i.d.	t.i.d.	t.i.d.	b.i.d.
Time from qualifying MI to entry						
mean	70 days	85 mo.	40 days	8 days	20 mo.	25 mo.
range	0,5-0,6 mo.	21 days-22yr	28-42 days	days-weeks	2-60 mo.	2-60 mo.
Duration of patients follow-up (months)						
mean	11,9	22,0	24,0	12,0	41,0	39,6
range	2-30	10-28	24-24	12-12	35-48	35-48

MI= myocardial infraction
o.d.: once daily
b.i.d.: twice daily
t.i.d: three times daily

Table 6. 4: Aspirin in coronary disease: design features of the studies

As we can see in Table 6.4, some of the studies include only males (studies 1 and 2) compared to others which include both males and females. Furthermore, in study 6 we have big daily dose (1000 mg) close enough to studies 2, 4 and 5. The mean duration is between 8 days and 85 months, while study 6 has duration 25 months. As we can see, there are some differences among the studies nothing though can cause such heterogeneity except maybe from the number of patients who participate in study 6.

The second thing that Canner checked was the adjustment of the individual study estimates for prognostic factors. From the examination of the seven risk factors[8] he found that the occurrence was significantly higher in the aspirin group than the placebo one. That finding was able to explain, at least in some level, the negative finding of the study. For the studies 2, 5 and 6 there was the possibility of adjusting the log-odds ratio estimate for some of the baseline

[8] The risk factors are: the history of congestive heart failure, the history of angina pectoris, the history of ECG- documented arrhythmia, the use of digitalis, the use of nitroglycerin, the use of propranolol or other beta-blockers and the use of other drugs.

characteristics but for all the other different baseline variables the adjustment was taken separately for every study.

The parameter of interest, θ, we use in this example is the log-odds ratio for mortality on aspirin compared to the placebo. So, the negative values show that aspirin is better than the control. Our interest focuses on the test of the null hypothesis that θ is equal to 0; against the alternative that θ is negative. The following Table 6.5 gives the one-sided p-values calculated from the Wald chi-squared statistic:

Study	Log-odds ratio	Std. error	Wald	p_{1i}	T_i
1	-0.329	0.197	2.78	0.048	6.09
2	0.385	0.203	3.59	0.029	7.08
3	0.216	0.275	0.62	0.216	3.07
4	0.22	0.143	2.35	0.063	5.54
5	0.0225	0.188	1.44	0.115	4.33
6	0.125	0.098	1.62	0.898	0.21
Total					26.32
X^2=26.32; (12df) p=0.01					

Table 6. 5: Fisher's combination of p-values

From the above table, we can see that the X^2 is 26.32 with 12 degrees of freedom. Note in this point that the degrees of freedom are two times the number of studies. Moreover, the fact that p=0.01 shows that there is statistically significance difference against the placebo. However, the disadvantage of the Fisher's method is that gives the same factor to all 6 studies, reducing the influence of the sixth study compared to the traditional meta-analysis presented in Table 6.3.

There are other methods related to Fisher's, like the one of Tippett (1931) [27] and Stouffer et al. (1949) [28]. Tippett's p test rejects the null hypothesis against the one sided alternative, i.e. that the two treatments are equivalent against that in at least one of the studies included the new treatment is

better than the control one [5]. That is happening if any of the p_{1i}, i=1,…,r is less than α^*, when:

$$\alpha^* = 1-(1-\alpha)^{1/r}$$

where α is the significance level for the combined significance test.

Stouffer et al. method uses the statistic:

$$U_r = \frac{\sum_{i=1}^{r} u(p_{1i})}{\sqrt{r}}$$

In the above formula we have that $u(p_{1i}) = \Phi^{-1}(1-p_{1i})$. When

$U_r > u(\alpha)$ then we reject the null hypothesis at level α and we accept the one-sided alternative.

However, both studies have the same disadvantage with Fisher's method, i.e. that every study gets the same weight instead of giving in every study a special factor. Mosteller and Bush (1954) [28] proposed a method which is a generalization of the Stouffer et al. in which each of the standard normal deviates $u(p_{1i})$ is weighted differently. We replace the U_r with U_{gr} and we get:

$$U_{gr} = \frac{\sum_{i=1}^{r} g_i u(p_{1i})}{\sqrt{\sum_{i=1}^{r} g_i^2}}$$

and

$$\sum_{i=1}^{r} g_i^2 = 1$$

The above method also referred to as the "weighted sum of zs method".

According to the first method, every weight g_i, i=1,…,r is equal to $1/\sqrt{r}$ and then $U_{gr} = U_r$. Suppose now that the same outcome measure has been recorded in each trial and that the parameter measuring treatment difference is

also identically defined [5]. If we calculate all the p_{1i} according to the assumption that:

$$\hat{\theta}_i \sim N(\theta, w_i^{-1})$$

Then $p_{1i} = 1 - \Phi(\hat{\theta}_i\sqrt{w_i})$ and $u(p_{1i}) = \hat{\theta}_i\sqrt{w_i}$. If we set $g_i = \sqrt{w_i}$ we get:

$$U_{gr} = \frac{\sum_{i=1}^{r} \hat{\theta}_i w_i}{\sqrt{\sum_{i=1}^{r} w_i}}$$

In this case U_{gr}^2 and the U statistic (formula (4.6)) are equal, i.e. it is equal to the test statistic for testing the treatment difference in the traditional fixed-effects meta-analysis.

In the case where we do not have any information on w_i it might be better to use $\sqrt{n_i}$ instead of the g_i we mentioned before, where n_i represents the total number of patients in the two treatment groups. If we use these weights in the example of the Canner, the statistic U_{gr}=1.65 with one- sided p-value equal to 0.049, as we can see in the following Table 6.5:

Study	p_{1i}	$u(p_{1i})$	$u(p_{1i})\sqrt{n_i}$	n_i
1	0.048	1.67	58.7	1239
2	0.029	1.9	74.11	1529
3	0.216	0.79	19.67	626
4	0.063	1.53	62.9	1682
5	0.115	1.2	41.91	1216
6	0.898	-1.27	-85.49	4524
Total			171.8	10816
U_{gr} =171.80/√10816=1.65;pi=0.049				

Table 6. 6: Mosteller and Bush method of combining p-values with weights equal to the square root of the study sample size, applied to the Canner data.

If we compare the results presented in the Table 6.6 with the initial results presents in the Table 6.3, we observe that those are much closer to the results from the fixed effects meta-analysis. In that point we must underlying that the p-value which is equal to 0.11 associated with U is a two-sided p-value, so the one-sided is 0.055. Moreover, since it is used the same outcome measure

in all trials, it is a reasonable approach to weight by the square root of the sample size, something that it may not be suitable for other data sets where there are different outcome measures.

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071

REFERENCES:

[1]: **Glass, G.V. (1976).** Primary, secondary and meta- analysis of research, *Educational Researcher*, 5, 3-8

[2]: **Sackett et al. (1997).** *Evidence- Based Medicine, How to Practice and Teach EBM*. Churchill-Livingstone, London

[3]: **Chalmers, T.C. and Lau, J. (1993).** Meta- analytic stimulus for changes in clinical trials, *Statistical Methods in Medical Research*, 2, 161-172

[4]: **Thompsnon, S. G. (1994).** Why sources of heterogeneity in meta-analysis should be investigated, *British Medical Journal*, 309, 1351-1355

[5]: **Whitehead, Anne (2002).** Meta- analysis of controlled clinical trials

[6]: **Normand, S.-L.T. (1999).** Tutorial in biostatistics: Meta-analysis: formulating, evaluating, combining and reporting, *Statistics in Medicine*, 18, 321-359

[7]: **Steward, L.A. and Clarke, M.J. (1995)** on behalf of the Cochrane working group on meta-analysis using individual patient data, Practical methodology of meta-analyses using updated individual patient data, *Statistics in Medicine*, 14, 2057-2079

[8]: **Bishop, Fienberg and Holland.** Discrete Multivariate Analysis

[9]: **International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (1998)**. ICH *Topic* E9: *Statistical Principles for Clinical Trials*

[10]: **Little, R.J.A. and Rubin, D.B. (1987).** *Statistical Analysis with Missing Data*. Wiley, New York

[11]: **Little, R. J. A. (1995).** Modelling the drop-out mechanism in longitudinal studies, *Journal of the American Statistical Association*, 90, 1112-1121

[12]: **Rubin, D. B. (1987).** *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York

[13]: **Greenland, S. and Salvan, A. (1990).** Bias in the one-step method for pooling study results, *Statistics in Medicine*, 9, 247-252

[14]: **Sprott, D.A (1973).** Normal likelihoods and their relation to large sample

theory of estimation, *Biometrika*, 60, 457-465

[15]: **Glass, G. V. (1976).** Primary, secondary and meta-analysis of research, *Educational Researcher*, 5, 3-8

[16]: **Hedges, L.V. and Olkin, J. (1985).** *Statistical Methods for Meta-Analysis*. FL: Academic Press, Orlando

[17]: **Cohran, W.G. (1954).** The combination of estimates from different experiments, *Biometrics*, 10, 101-129

[18]: **DerSmonian, R. and Laird, N. (1986).** Meta-analysis in clinical trials, *Controlled Clinical Trials*, 7, 177-188

[19]: **Whitehead, A., Bailey, A. and Elbourne, D. (1999).** Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis, *Journal of Biopharmaceutical Statistics*, 9, 1-16

[20]: **Fisher, R.A. (1932).** Statistical Methods for Research Workers (4th edn). Oliver and Boyd, London

[21]: **Senn, S. (2000).** The many modes of meta, *Drug Information Journal*, 34, 535-549

[22]: **Hardy, R. J. and Thompson, S. G. (1998).** Detecting and describing heterogeneity in meta- analysis, *Statistics in Medicine*, 17, 841-856

[23]: **Qizilbash, N., Whitehead, A., Higgins, J., Wilcock, G., Schneider, L. and Farlow, M. (1998)** on behalf of the Dementia Trialists' Collaboration Cholinesterase inhibition for Alzheimer disease: a meta-analysis of the tacrine trials

[24]: **Jones, D. R. and Whitehead, J. (1979).** Sequential forms of the log rank and modified Wilcoxon tests for censored data, *Biometrica*, 66, 105-113. Correction (1981), *Biometrica*, 68, 576

[25]: **Canner, P.L. (1987).** An overview of six clinical trials of aspirin in coronary heart disease, *Statistics in Medicine*, 6, 255-263

[26]: **Tippet, L.H.C. (1931).** The Methods of Statistics. Williams and Norgate, London

[27]: **Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. and Williams, R.M., Jr (1949).** The American Soldier: Adjustment during

Army Life, Vol. 1. Princeton, NJ: Princeton University Press

[28]: **Mosteller, F. and Bush, R.R. (1954).** Selected quantitative techniques, in G.Lindsey (ed), *Handbook of Social Psychology:* Vol.1, *Theory and Method*. MA: Addison-Wesley, Cambridge

https://doi.org/10.26219/heal.aueb.6071

https://doi.org/10.26219/heal.aueb.6071