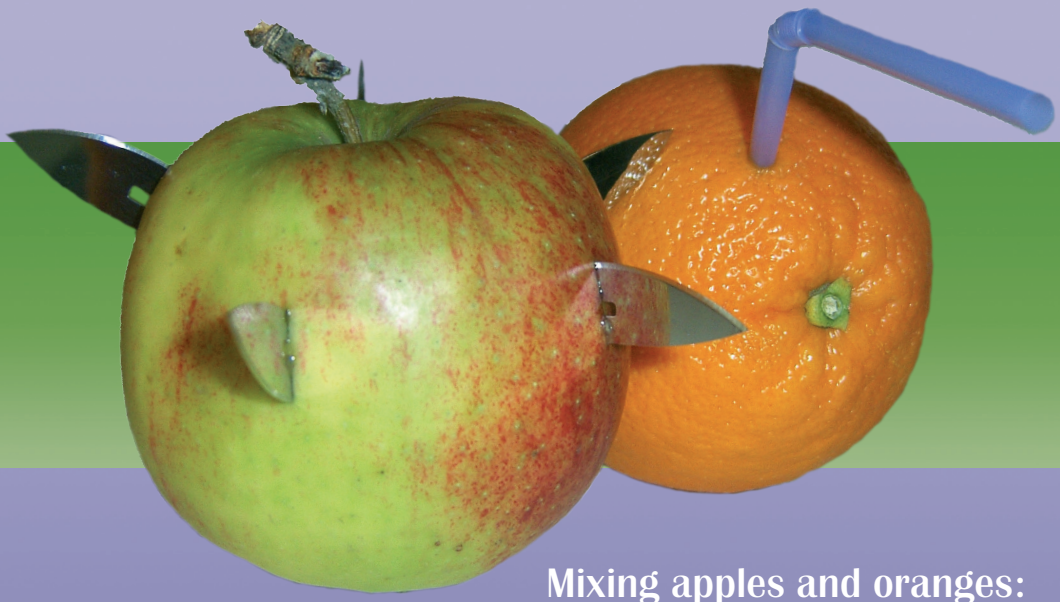
Multivariate meta-analysis: modelling the heterogeneity



Mixing apples and oranges: dangerous or delicious?

Lidia R. Arends

MULTIVARIATE META-ANALYSIS: MODELLING THE HETEROGENEITY

Mixing apples and oranges: dangerous or delicious?

Lidia R. Arends

Acknowledgements

The publication of this thesis was financially supported by:
The Department of Epidemiology & Biostatistics of the Erasmus MC, Erasmus University Rotterdam, GlaxoSmithKline, Serono Benelux BV, Boehringer Ingelheim BV, Pfizer BV and the Dutch Cochrane Centre.

Cover design: Bureau Stijlzorg, Utrecht
Layout: EM Osseweijer, Etten-Leur
Printed by: Haveka BV, Alblasserdam

ISBN 90-9020786-4

Multivariate Meta-analysis: Modelling the Heterogeneity

Mixing apples and oranges: dangerous or delicious?

Multivariate meta-analyse: het modelleren van de heterogeniteit

Het mengen van appels en peren: gevaarlijk of heerlijk?

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 28 juni 2006 om 15.45 uur

door

Lidia Roelfina Arends
geboren te Eelde

Promotiecommissie

Promotor:	Prof.dr. Th. Stijnen
Overige leden:	Prof.dr. M.G.M. Hunink
	Prof.dr. J.C. van Houwelingen
	Prof.dr. J.D.F. Habbema

Contents

Manuscripts based on studies described in this thesis

Chapter 2

Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T.
Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses.
Statistics in Medicine 2000; 19(24): 3497-3518.

Chapter 3

van Houwelingen HC, Arends LR, Stijnen T.
Advanced methods in meta-analysis: multivariate approach and meta-regression.
Statistics in Medicine 2002; 21(4): 589-624.

Chapter 4

Arends LR, Voko Z, Stijnen T.
Combining multiple outcome measures in a meta-analysis: an application.
Statistics in Medicine 2003; 22(8): 1335-1353.

Chapter 5

Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T.
Multivariate random-effects meta-analysis of ROC curves.
Medical Decision Making. Provisionally accepted.

Chapter 6

Arends LR, Hunink MGM, Stijnen T.
Meta-analysis of summary survival curve data.
To be submitted.

"Of course it mixes apples and oranges; in the study of fruit nothing else is sensible;
comparing apples and oranges is the only endeavor worthy of true scientists;
comparing apples to apples is trivial."

Gene V. Glass, 2000

1

Introduction

1 Introduction

This thesis is about multivariate random effects meta-analysis and meta-regression. In this introduction these terms will be explained and an outline of the thesis will be given.

1.1 What is meta-analysis?

Meta-analysis may be broadly defined as the quantitative review and synthesis of the results of related but independent studies[1]. These studies usually originate from the published literature. For the purpose of critically evaluating a clinical hypothesis based on published clinical trials, meta-analysis is an efficient tool for summarizing the results in the literature in a quantitative way. In most of the cases it results in a combined estimate and a confidence interval[2]. Meta-analysis allows for an objective appraisal of the evidence, which may lead to resolution of uncertainty and disagreement. It can reduce the probability of false-negative results and thus prevent undue delays in the introduction of effective treatments into clinical practice. A priori hypotheses regarding treatment effects in subgroups of patients may be tested with meta-analysis[3] as well. It may also explore and sometimes explain the heterogeneity between study results, see the section on meta-regression below.
Since the introduction in 1976[4] of the term 'meta-analysis', it has become an increasingly important technique in medical research. This is illustrated in Figure 1, where the number of studies found in Medline containing the keyword 'meta-analysis' is plotted against the year of publication.

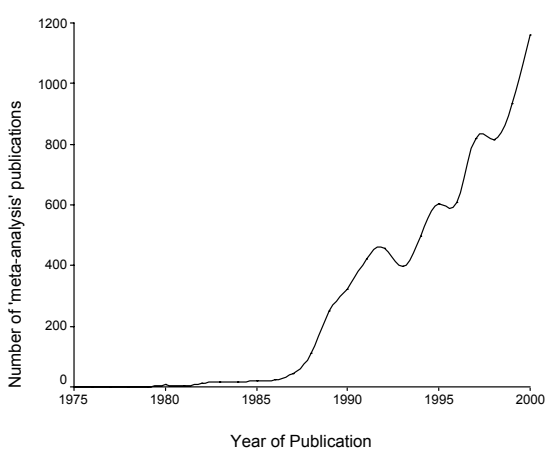


Figure 1. The meta-analysis trend

With the increasing popularity of meta-analysis, also the field of application of statistical meta-analysis methods is growing. In earlier days the main and often only interest was to statistically pool the results of independent but 'combinable' studies[5] to increase power, resulting in an overall estimate of one specific outcome measure and a confidence interval. For this situation most meta-analysts know how they can analyse the collected data. Nowadays we are often faced with meta-analysis of more complex medical data, and there are many practical situations where appropriate statistical meta-analytic methods are still lacking or underdeveloped. New statistical methods are needed to meta-analyse these complex data types.

1.2 Fixed and random effects meta-analysis

In every meta-analysis the point estimates of the effect size will differ between the different studies in the meta-analysis, at least to some degree. One cause of these differences is sampling error, which is present in every estimate. When observed effect sizes differ only due to sampling error, the true underlying study specific effects are called homogeneous. In this case the differences between the estimates are just random variation, and not due to systematic differences between studies. In other words, the true underlying effect size is exactly the same in each study. In that case, if every study would be infinitely large, all studies would yield an identical result. The case of homogeneity can be accommodated in meta-analysis by using a what is called the 'fixed effects model'[6]. In the early days of meta-analysis statistical modelling was always done under the assumption that the true effect measure was homogeneous across all studies, thus with a fixed effects model.

However, often the variability in the effect size estimates exceeds that expected from sampling error alone, i.e. there is not just one and the same true effect for each study, but 'real' differences exist between studies. In this case we say that there is heterogeneity between the treatment effects in the different studies. In a famous paper DerSimonian and Laird (1986)[7] introduced a statistical model that allows heterogeneity in the true treatment effects. In that model the different true study specific effects are assumed to have a distribution. This distribution is characterized by two parameters, the mean and the standard deviation, and both have to be estimated from the data. The first is the parameter of main interest, and is interpreted as the average effect. The other parameter is called the between studies standard deviation and describes the heterogeneity between the true effects. This model is called the 'random effects model'. This model is tending to become the standard method for the simple case where the meta-analysis is focused on a single (univariate) effect measure, e.g. one treatment effect. See for instance the review article of Normand (2000)[1].

The fixed effects method to estimate a common treatment effect yields a narrower confidence interval than the random effects estimation of an average treatment effect when there is heterogeneity observed between the results of the different trials[8]. This explains why the fixed effects method is still often used. The simplistic assumption of a common treatment effect in all trials used in a fixed effects analysis ignores the potential between-trial component of variability and can lead to over dogmatic interpretation[9]. Since the trials in a meta-analysis are almost always clinically heterogeneous, it is to be anticipated that to some extent their quantitative results will be statistically heterogeneous[10]. Hence a random effects model appears more justified than a fixed effects model.

1.3 Meta-regression: Is there an explanation for heterogeneity?

In the previous section we discussed the term heterogeneity, i.e. the part of the variability in the outcome measure across studies not due to within study sampling variability. If there is much heterogeneity between the studies, one could question whether it is wise to combine the studies at all. However, heterogeneity can be regarded as an asset rather than a problem. It allows clinically and scientifically more useful approaches attempting to investigate how potential sources of heterogeneity impact on the overall treatment effect[10]. For example, the treatment effect could be higher in trials that included a lot of old males, whereas the treatment effect could be lower in studies with a lot of young female patients. The dependence of the treatment effect on one or more characteristics like mean age and sex of the trials in the meta-analysis can be explored via meta-regression. In meta-regression the trial characteristics are put as covariates in a regression analysis with the estimated treatment effect of the trial as dependent variable. Ideally the covariates used in such analyses should be specified in advance to reduce the risk of post hoc conclusions prompted by inspecting the available data[8]. Otherwise there is a danger of false positive results. This is in particular the case when a fixed effects regression model is used. For example, consider the case of just two studies producing estimates with non-overlapping confidence intervals. Any covariate of which the value differs between the studies will be significantly related to the heterogeneity among the studies, and hence a potential explanation of it. It is clear, however, that the majority of such 'explanations' will be entirely spurious.[11]. As the number of studies increases, the risk of identifying spurious associations decreases as long as there is only a limited number of covariates.

The statistical purpose of meta-regression is to see to what extent covariates can explain the between-trial component of the variance. In case all between-trial variation is explained by the covariates, the random effects meta-regression reduces

to a fixed effects regression model, in which al variability is explained as sampling variability. Using covariates to explain the differences in treatment effect across the trials could lead to a better scientific understanding of the data and more clinically useful conclusions on which to base decisions about medical interventions[8, 10].

1.4 Univariate versus multivariate models

In a meta-analysis clinical interest does not always concern only one specific outcome measure. Sometimes the focus is on the combination of several outcome measures that are presented in the individual studies, for instance when there are more treatment groups or more outcome variables. When the summary data per study are multi-dimensional, the data analysis is unfortunately usually restricted to a number of separate univariate analyses, i.e. one analysis per outcome variable. However, such univariate analyses neglect the relationships between the multiple outcome measures. In a multivariate analysis all outcome measures are analysed jointly, therefore also revealing information about the correlations between the multiple outcome variables.

1.5 Aim and outline of this thesis

The aim of this thesis is to develop new statistical methods and improve existing ones for the analysis of meta-analysis data from medical studies. We will specifically focus on multivariate random effects meta-analysis approaches.

As first topic the relationship between baseline risk (i.e. the risk in the control group) and treatment effect is investigated as a possible explanation of between-study heterogeneity in clinical trial meta-analysis. This is a very special case of meta-regression. The standard approach is seriously flawed for several reasons and can lead to very misleading results[12]. In chapter 2 a Bayesian approach to the problem is proposed, based on a bivariate meta-analysis model. Different from other proposed methods, it uses the exact rather that an approximate likelihood. Besides it explicitly models the distribution of the underlying baseline risks, in contrast to the method of Thompson et al. (1997)[13].

In chapter 3 advanced statistical methods for meta-analysis are reviewed such as bivariate meta-analysis[14] and meta-regression. It is shown that these methods fit into the framework of the general linear mixed model. Next to discussing the underlying theory, much attention is given to how these analyses can be carried out using the mixed model procedures of standard statistical packages.

In chapter 4 a meta-analysis is considered to the effect of surgery compared to conservative treatment in patients with increased risk of stroke on stroke-free survival. This is the first published meta-analysis in which more than two outcome measures are simultaneously analyzed using the multivariate meta-analysis model. It

is shown that a multivariate analysis can reveal substantially more information than separate univariate analyses.
In chapter 5 the statistical meta-analysis is considered of ROC curve data where each study contributes one pair of specificity and specificity. Until now, this type of data has been analyzed with rather ad hoc approaches. In this chapter it is shown that this type of data nicely fits into the bivariate meta-analysis framework. A random intercept model is fitted with approximate and exact likelihood. Moreover the model is extended with a random slope next to a random intercept.
In chapter 6 a multivariate random effect model is proposed for the joint analysis of survival proportions reported at multiple times in different studies. The model can be seen as a generalization of the fixed effect model of Dear[15] and is illustrated with a simulated as well as with a clinical data example.
The methods discussed in this thesis have their specific benefits and limitations. In chapter 7, the general discussion, these will be put in perspective. Furthermore, some directions are given for future research.

References

1. Normand S-L. Meta-analysis: formulating, evaluating, combining and reporting. *Statistics in Medicine* 1999; **18**:321-359.
2. van Houwelingen JC. Meta-analysis; Methods, limitations and applications. *Biocybernetics and Biomedical Engineering* 1995; **15**(1-2):53-61.
3. Zou K. What is evidence-based medicine? *Academic Radiology* 2004; **11**(2):127-133.
4. Glass GV. Primary, secondary and meta-analysis of research. *Educational Researcher* 1976; **5**: 3-8.
5. Egger M, Ebrahim S, Smith GD. Editorial: Where now for meta-analysis? *International Journal of Epidemiology* 2002; **31**(1):1-5.
6. Sutton A, Abrams K, Jones D, Sheldon T, Song F. *Methods for meta-analysis in medical research*. Wiley: Chichester, 2000.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177-188.
8. Armitage P, Colton T. *Encyclopedia of Biostatistics (1st ed., Vol. 4)*. John Wiley & Sons, Inc.: New York, 1998.
9. Thompson S, Pocock S. Can meta-analysis be trusted? *Lancet* 1991; **338**:1127-1130.
10. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**(6965):1351-1355.
11. Higgins J, Thompson S. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* 2004; **23**:1663-1682.
12. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735-738.
13. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741-2758.
14. van Houwelingen H, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2272-2284.
15. Dear KBG. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994; **50**:989-2001.

2

Baseline risk as predictor of treatment benefit

Abstract

A relationship between baseline risk and treatment effect is increasingly investigated as a possible explanation of between-study heterogeneity in clinical trial meta-analysis. An approach that is still often applied in the medical literature, is to plot the estimated treatment effects against the estimated measures of risk in the control groups (as a measure of baseline risk), and to compute the ordinary weighted least squares regression line. However, it has been pointed out by several authors that this approach can be seriously flawed. The main problem is that the observed treatment effect and baseline risk measures should be viewed as estimates rather than the true values. In recent years several methods have been proposed in the statistical literature to potentially deal with the measurement errors in the estimates. In this chapter we propose a vague priors Bayesian solution to the problem which can be carried out using the 'Bayesian inference Using Gibbs Sampling' (BUGS) implementation of Markov chain Monte Carlo numerical integration techniques. Different from other proposed methods, it uses the exact rather than an approximate likelihood, while it can handle many different treatment effect measures and baseline risk measures. The method differs from a recently proposed Bayesian method in that it explicitly models the distribution of the underlying baseline risks. We apply the method to three meta-analyses published in the medical literature and compare the results with the outcomes of the other recently proposed methods. In particular we compare our approach to McIntosh's method, for which we show how it can be carried out using standard statistical software. We conclude that our proposed method offers a very general and flexible solution to the problem, which can be carried out relatively easily with existing Bayesian analysis software. A confidence band for the underlying relationship between true effect measure and baseline risk and a confidence interval for the value of the baseline risk measure for which there is no treatment effect are easily obtained by-products of our approach.

1 Introduction

Meta-analysis is an increasingly common type of analysis for combining the results from several clinical trials to obtain an overall assessment of treatment effectiveness of a certain medical intervention. A common criticism about meta-analyses is that they combine information from trials with very different patient characteristics and designs. These trial variations can result in much bigger differences in treatment outcome across the trials than one would expect on the basis of the sampling variability of the estimated treatment effects of the individual trials. Furthermore, the frequently used tests of homogeneity have little power, so when this test is not significant, there can still be heterogeneity that should not be ignored. An analysis which ignores the 'extra' heterogeneity in treatment outcome can be clinically misleading and scientifically naive[1-3]. Therefore, it is necessary to assess whether the heterogeneity in treatment effect can be explained by trial-level characteristics. For example, trials that included older (often higher-risk) patients could on average show a larger treatment effect than trials with younger patients. In clinical practice it can be important to know whether and how the expected treatment benefit varies according to certain patient characteristics, in order to assess the treatment effect in the clinician's own, specific population of patients[1] and to specify more specific therapeutic recommendations[3].

In this chapter we consider meta-analyses of clinical trials having one experimental and a control group. The outcome of interest is the occurrence of some specified clinical event, for example death or a certain disease. We assume to have data available only on trial-level and not on individual patients. Thus it is impossible to individually relate age or sex of a patient to the treatment outcome of that individual patient. In fact, the best one can do in such a situation is to relate aggregated information on trial level (such as mean age or percentage males) with the estimated treatment effect across the trials.

Brand and Kragt[1] were among the first who argued the importance of investigating the possible relationship between treatment effect and covariates. In particular they stimulated debate about the importance of considering the covariate 'baseline risk'. This baseline risk reflects the risk of the outcome event for a patient under the control condition and indicates the average risk of a patient in that trial if he or she was not treated. Heterogeneity in baseline risk among the trials is likely to reflect differences in patient characteristics like age, medical history and comorbidity of the patient populations included in the several trials and might therefore result in different treatment effects among the trials[4]. In other words, the baseline risk of the study population may markedly modify the absolute effect of the intervention in a given

trial. In some recent meta-analyses this relation between treatment effect and baseline risk was even one of the central issues, with the aim to define which patients would benefit most and which least from the medical intervention[5-12].

As a graphical illustration of the relationship between treatment effect and baseline risk, one usually plots the estimated treatment effects against the observed percentage of events in the corresponding control groups. A straightforward, but problematic, way of assessing the possible relationship is to compute the ordinary weighted least squares (WLS) regression line. In the medical literature, this method is still mostly applied. However, this conventional method has potential pitfalls and has been seriously criticised[4, 13-18]. The main problem is that the observed baseline risk and the observed treatment effect in a trial are estimated from a finite sample, and therefore are estimates rather than true values[4]. Consequently, one should account for the measurement errors in these variables. If not, regression to the mean[13] and attenuation due to measurement errors[19] could seriously bias the slope of the regression line of treatment effect versus baseline risk. For an overview of the approaches followed in practice and the associated statistical problems, the reader is referred to Sharp et al.[16].

In the recent statistical literature methods are proposed that account for the above mentioned problems[4, 17, 18]. McIntosh[4] assumed a bivariate normal distribution for the underlying true treatment effect and the true baseline risk measure, together with an approximate normal measurement errors model. The model was fitted with standard likelihood or Bayesian methods. Van Houwelingen et al.[20] assumed the same model, but did not mention that the method could be directly used to estimate the relationship between true treatment effect and true baseline risk. Walter[18] assumed a linear functional relationship between true treatment effect and true baseline risk (i.e. a model without residual variation) and an approximate normal measurement errors model. Standard likelihood methods were used to fit the model.

In this chapter we propose a hierarchical Bayesian modelling approach. Our method differs from McIntosh's method in that it uses an exact measurement errors model. In addition we also generalise the assumed bivariate normal distribution for the true treatment effect and true baseline risk to a mixture of two bivariate normal distributions. We fit our models following a vague priors Bayesian approach using the BUGS[21] implementation of Gibbs sampling. Our approach is in the spirit of that of Thompson et al.[17]. The difference with their method is that we explicitly model the distribution of the true baseline risk measures.

In this chapter we will re-analyse three meta-analyses published in the medical literature, and compare our method with some of the other recently proposed methods and the WLS approach. In section 2 we introduce the three meta-analysis

examples and indicate the problems of the conventional WLS method. In section 3 our modelling approach is presented and differences with other recently proposed methods for this problem are described. In section 4 the method of estimation is specified and in section 5 the results of the proposed methods are given and compared with the WLS approach and with the methods of McIntosh[4] and Thompson et al.[17]. Finally, we conclude with a discussion in section 6. Appendix 1 contains the BUGS code needed to perform the analyses with our modelling approach, and Appendix 2 contains a SAS program to do some of the analyses using approximate likelihood.

2 Examples

In this section we introduce three motivating examples of meta-analyses from the medical literature, in which the possible relationship between treatment effect and baseline risk was a central issue. The problems with the standard statistical analysis are briefly pointed out. In section 5 the results of the re-analyses of these three studies are given.

2.1 Meta-analysis example 1: Effect of tocolysis therapy on pre-term birth

Brand and Kragt[1] were among the first authors who argued that presenting one pooled odds ratio as 'the' treatment effect in a meta-analysis can be misleading if the odds ratio depends on the baseline risk. They reported on a meta-analysis of 14 placebo-controlled trials[22] evaluating the effect of tocolysis with β-mimetics to delay pre-term deliveries in high risk mothers. The treatment effect was measured as the (log) odds ratio of pre-term birth in the treatment group relative to the control group. The data are shown in Table 1. The research question was whether the treatment effect depends on the proportion of pre-term births in the control group. The proportion of events in the control group served as the measure of baseline risk, indicating the
risk of an average patient in a trial if no treatment was applied. Large differences among the trials in this respect may reflect for instance different selection criteria between them and hence a constant odds ratio over trials is not to be expected. After ordering the trials according to increasing proportion of pre-term deliveries in the control group, the trend in the odds ratio was striking (Table 1).

Table 1. Number of deaths, total number of persons and corresponding risks of pre-term births in the treatment and control group of the randomised trials in the meta-analysis of Brand & Kragt[1] (ordered to increasing baseline risk).

Source	Treatment group		Control group		Odds Ratio
	preterm births / number of women	risk of pre-term birth	preterm births / number of women	risk of pre-term birth (baseline risk)	
Mariona	0 /4	0%	0 /5	**10%**	**1.00**
Howard et al.	2 /15	13%	2 /18	**11%**	**1.22**
Larsen et al.	11 /131	8%	6 /45	**13%**	**0.57**
Hobel	2 /16	13%	3 /15	**20%**	**0.58**
Calder et al.	4 /37	11%	9 /39	**23%**	**0.43**
Scommegna	1 /15	7%	5 /17	**29%**	**0.24**
Larsen et al.	5 /49	10%	16 /50	**32%**	**0.27**
Christensen et al.	0 /14	0%	6 /16	**38%**	**0.10**
Leveno et al.	15 /54	28%	25 /52	**48%**	**0.42**
Wesselius et al.	6 /33	18%	15 /30	**50%**	**0.24**
Cotton et al.	6 /19	32%	11 /19	**58%**	**0.35**
Barden	0 /12	0%	8 /13	**62%**	**0.07**
Ingemarsson	0 /15	0%	10 /15	**67%**	**0.06**
Spellacy et al.	6 /14	43%	11 /15	**73%**	**0.30**
Overall Odds ratio					**0.30**

Figure 1 shows the plot of the observed log odds ratio against the observed proportion of pre-term deliveries in the placebo group. The slope of the ordinary weighted least squares (WLS) regression line turned out to be statistically significantly negative (p=0.03), suggesting a better treatment effect with increasing baseline risk.

Several problems may arise with this approach. The main problem, pointed out by Senn[13], is that in the regression the dependent variable 'treatment effect' includes the independent variable 'baseline risk', which causes a functional relationship between the dependent and the independent variable.

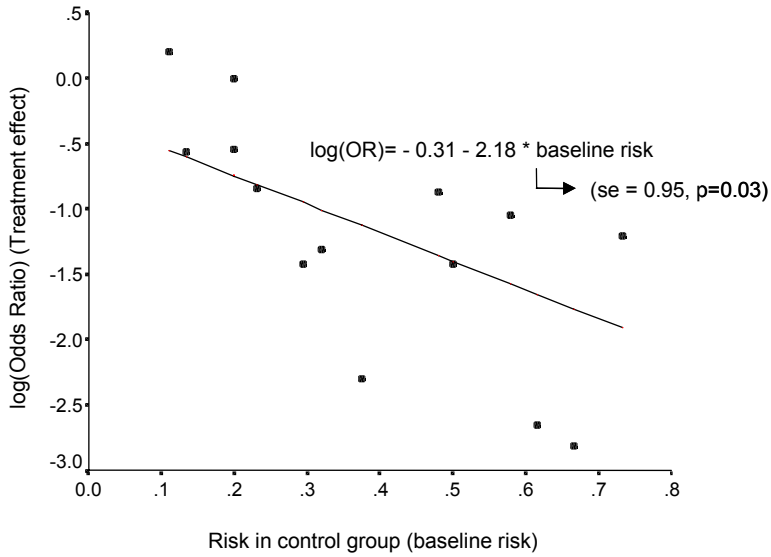


Figure 1. Original WLS regression analysis as published in the meta-analysis of Brand & Kragt[1]

A positive error in the observed baseline risk is associated with a negative error in the estimated log odds ratio, leading to a negative bias in the estimated slope or a spurious negative correlation. A thorough discussion of the problems associated with the WLS approach is given by Sharp et al.[16]. They discuss the statistical pitfalls of this approach at the hand of the following three types of graphs currently encountered in the medical literature.

A. Plot of the treatment effect against the proportion of events in the control group.
B. Plot of the treatment effect against the average proportion of events in the control and treatment group.
C. Plot the proportion of events in the treated group against the proportion of events in the control group (see example 2).

In all cases the WLS approach leads to misleading conclusions. As explained above, WLS in graph type A leads to bias due to regression to the mean. The type B graph is sometimes used to avoid this problem. However, as explained by Sharp et al.[16], WLS in type B graphs is only valid under the assumption of no variation in true treatment effects, a strange assumption in this context. Apart from that, the interpretation of type B graphs is problematic. For all graph types, since the proportion of events in the control group is measured with error, the WLS method suffers from bias towards zero in the slope of the regression line. Even if the dependent and independent variable were not functionally related, the estimated

slope could be biased towards zero by the fact that the dependent variable is measured with error. A minor problem in all graphs is that the weights used in the WLS approach only reflect sampling error, and do not account for residual heterogeneity between trials with the same baseline risk. In general the biases associated with the WLS approach will be particular acute in meta-analyses which include some small trials or in which the true variability in underlying risks across trials is small.

2.2 Meta-analysis example 2: Drug treatment in mild-to-moderate hypertension

In 1995 Hoes et al.[6, 23] published a meta-analysis of clinical trials in which drug treatment was compared to placebo or no treatment with respect to (cardiovascular) mortality in middle-aged patients with mild-to-moderate hypertension. Twelve trials, which showed considerable variation in the risk of mortality in the control groups, were included in the meta-analysis. Unlike to the previous meta-analysis, the data are presented as the number of events and the (partially estimated) total number of person-years per group instead of the number of events and the sample size. The data are given in Table 2.

The research question in this meta-analysis was whether drug treatment prevents death in mild-to-moderate hypertensive patients and whether the size of the treatment effect depends on the event rate in the control group (baseline rate). To avoid the functional relationship between the variables at the dependent and independent variable, as was the case in the previous meta-analysis, a 'l'Abbé plot' was presented. In this l'Abbé plot[24], shown in Figure 2, the observed mortality rate per 1000 person-years in the treatment groups is plotted against the observed mortality rate per 1000 person-years in the control groups. The dotted line of identity corresponds with no treatment effect. For trials below this line the observed death rate in the treatment group is lower than in the control group, suggesting a beneficial effect of drug treatment. On the other hand, for trials falling above this line, drug treatment seems to unfavourably influence mortality.

To study the relationship between the mortality rates in the treatment versus control group, the WLS regression line was determined, represented by the solid line in Figure 2. The WLS regression line has a slope statistically significantly smaller than one and intersects the no-effect line for a positive value of the baseline mortality rate. In the sequel the baseline risk corresponding with no effect is called the break-even point. Its estimate in this meta-analysis is about 6 per 1000 person-years. The authors conclude that drug treatment may reduce mortality when treatment is initiated in those beyond this break-even point. At a lower mortality rate than about 6 per 1000 person years, treatment has no influence or may even increase mortality.

Table 2. Number of deaths, total number of person years and corresponding incidence rates of all-cause mortality in the treatment and control group of the randomised trials in mild-to-moderate hypertension in the meta-analysis of Hoes et al.[6].

Source	**Treatment group**			**Control group**		
	deaths /	#person-years	mortality rate /1000prsy	deaths /	#person-years	mortality rate /1000prsy **(baseline rate)**
VA	10 /	595.2	16.8	21 /	640.2	32.8
VA-NHBLI	2 /	762.0	2.6	0 /	756.0	0.0
HDFP	54 /	5635.0	9.6	70 /	5600.0	12.5
HDFP	47 /	5135.0	9.2	63 /	4960.0	12.7
HDFP	53 /	3760.0	14.1	62 /	4210.0	14.7
Oslo	10 /	2233.0	4.5	9 /	2084.5	4.3
ANBPS	25 /	7056.1	3.6	35 /	6824.0	5.1
MRFIT	47 /	8099.0	5.8	31 /	8267.0	3.7
MRFIT	43 /	5810.0	7.4	39 /	5922.0	6.6
MRFIT	25 /	5397.0	4.6	45 /	5173.0	8.7
MRC men	157 /	22162.7	7.1	182 /	22172.5	8.2
MRC women	92 /	20885.0	4.4	72 /	20645.0	3.5

Since in the l'Abbé-plot the dependent and independent variable are not functionally related, the WLS analysis in this example does not suffer from the main problem raised to the WLS analysis in the previous example. Nevertheless the conclusions of this meta-analysis were heavily criticised with respect to the statistical method used. In an accompanying commentary Egger and Davey Smith[14] argued that it was a misleading analysis. The problem raised was that the slope of the regression line is biased towards zero when the incidence rates in the control groups are measured with error. This is the well known phenomenon of attenuation of the regression line due to measurement error, sometimes called regression dilution bias[25].

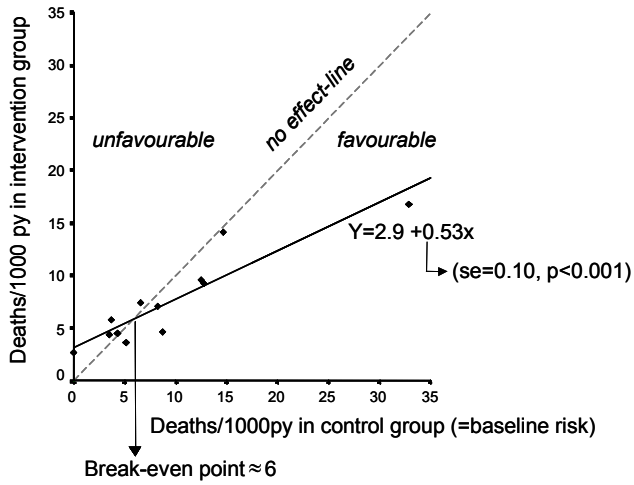


Figure 2. Original WLS regression analysis as published in the meta-analysis of Hoes et al.[6]

2.3 Meta-analysis example 3: Cholesterol lowering and mortality

The objective in our third meta-analysis example, published in 1993 by Davey Smith et al.[7], was to investigate the level of risk of death from coronary heart disease (CHD) above which cholesterol lowering treatment produces net benefits. This meta-analysis comprised 33 trials. Data are given in Table 3.

In Figure 3 the observed log odds ratio for total mortality is plotted against the observed rate of coronary heart disease (CHD). The zero level horizontal line corresponds with no treatment effect. The WLS regression analysis carried out by the authors shows a significant trend of increasing treatment benefit with increasing baseline CHD mortality risk (p<0.001). The authors concluded that currently evaluated cholesterol lowering drugs seem to produce total mortality benefits in only a small proportion of patients having a very high risk of death from CHD, namely those patients with a baseline risk larger than the break-even point, i.e. the intersection of the regression line with the zero level no effect line.

In this analysis, like in our first example, the dependent and independent variable are functionally related, causing a negative bias in the slope of the regression line. The authors were aware of this problem and, in order to circumvent it, they also performed an analysis with the observed total CHD mortality in treatment and control group together as independent variable.

Table 3. Number of deaths, total number of person years and corresponding incidence rates of all-cause mortality in the treatment and control group of the randomised trials in the meta-analysis of Davey Smith et al.[7].

Source	Treatment group			Control group		
	# deaths	# person-years	mortality rate / 1000 prsy	number of cases	# person-years	mortality rate / 1000 prsy **(baseline rate)**
Singh	28 /	380	73.68	51 /	350	145.71
Marmorston	70 /	1250	111.11	38 /	640	118.75
Stamler	37 /	690	53.62	40 /	500	80.00
McCaughan	2 /	90	22.22	3 /	30	100.00
Harrold	0 /	30	.00	3 /	30	100.00
Stockholm	61 /	1240	49.19	82 /	1180	69.49
Oslo Diet	41 /	930	44.09	55 /	890	61.80
Low Fat	20 /	340	58.82	24 /	350	68.57
DART	111 /	1930	57.51	113 /	1920	58.85
VA drug	81 /	1240	65.32	27 /	410	65.85
Newcastle	31 /	1140	27.19	51 /	1140	44.74
Oliver	17 /	210	80.95	12 /	220	54.55
Acheson	23 /	210	109.52	20 /	230	86.96
STARS	0 /	90	.00	4 /	170	23.53
CDP	1450 /	38620	37.55	723 /	19420	37.23
Dayton	174 /	1350	128.89	178 /	1330	133.83
Soya Bean	28 /	890	31.46	31 /	860	36.05
Scottish	42 /	1970	21.32	48 /	2060	23.30
Sahni	4 /	150	26.67	5 /	150	33.33
Upjohn	37 /	2150	17.21	48 /	2100	22.86
Sydney	39 /	1010	38.61	28 /	1120	25.00
Rose	8 /	100	80.00	1 /	50	20.00
NHLIB	5 /	340	14.71	7 /	340	20.59
Minnesota	269 /	4410	61.00	248 /	4390	56.49
POSCH	49 /	3850	12.73	62 /	3740	16.58
CLAS	0 /	190	.00	1 /	190	5.26
Frick '93	19 /	1510	12.58	12 /	1560	7.69
LCCPPT	68 /	13850	4.91	71 /	13800	5.14
Frick '87	46 /	10140	4.54	43 /	10040	4.28
EXCEL	33 /	5910	5.58	3 /	1500	2.00
WHO	236 /	27630	8.54	181 /	27590	6.56
SCOR	0 /	100	.00	1 /	100	10.00
Gross	1 /	20	50.00	2 /	30	66.67

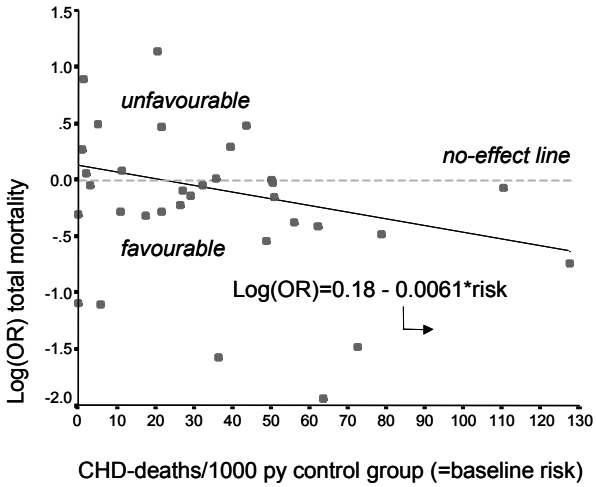


Figure 3. Original WLS regression analysis as published in the meta-analysis of Davey Smith et al.[7]

Since the results did not substantially change, they concluded that in this case the bias was negligible. Although taking total observed risk as independent variable repairs the problem of induced correlation, it cannot be regarded as a general solution since it leads to a regression analysis that is difficult to interpret[16]. Moreover, still the problem remains that the independent variable is measured with error[17].

3 Models

In this chapter we propose to model the data in a hierarchical way that explicitly makes a distinction between the regression model for the true effect measure on the true baseline risk measure, and the measurement error model. We denote the true underlying baseline risk measure of the i^{th} trial by ξ_i. This is for instance the true log odds of the event in the control group, as in example 1, or the true log event rate in the control group, as in example 2. The true measure to be related with ξ_i is denoted by η_i. This η_i could be a treatment effect measure, for example the true log odds ratio as in example 1, or a measure for the risk in the treatment group, for instance the true log event rate in the treated group, as in example 2. The estimates of ξ_i and η_i are denoted by $\hat{\xi}_i$ and $\hat{\eta}_i$.

The data are modelled following a hierarchical modelling approach, distinguishing the following three model components.

1. *Underlying regression model*
   Model for the regression of the true treatment effect measure (or risk measure in the treated group) η_i on the true baseline measure ξ_i:
   $\eta_i = \alpha + \beta\xi_i + \varepsilon_i \quad$ with $\quad \varepsilon_i \cong \mathrm{N}(0, \tau^2)$ (1)
   The residual variance τ^2 describes the heterogeneity in true treatment effects (or true risks under treatment) in populations with the same true baseline risk.

2. *Baseline risks model*
   Model for the distribution of the true baseline risk measures ξ_i:
   $\xi_i \cong G$ for some parametric model G

3. *Measurement errors model*
   Model for the 'measurement errors':
   $(\hat{\xi}_i, \hat{\eta}_i)$ given $(\xi_i, \eta_i) \cong F$ for some parametric model F

Notice that the first two components, called by McIntosh[4] the 'structural model', determine the joint distribution of ξ_i and η_i. McIntosh[4], Thompson et al.[17], and implicitly van Houwelingen et al.[20] assume the same underlying risk model as above. Walter[18] also assumes a linear relation between ξ_i and η_i as in our underlying regression model, but in his model the between trial variability τ^2 is set to zero. This means that it is assumed that there is no residual heterogeneity in treatment effects, or that all heterogeneity in treatment effect is explained by differences in baseline risk. See also the comment of Bernsen et al.[26].

A typical assumption for the distribution of the true baseline risks in component 2 of the model is a normal distribution:

$$\xi_i \cong \mathrm{N}(\bar{\xi}, \sigma_\xi^2) \quad (2)$$

This was the assumption made by McIntosh[4] and van Houwelingen et al.[20]. However, this seems to be a rather strong assumption[16] and there is no clear rationale for it. For instance, one could easily imagine that the distribution of the baseline risks would be bimodal, a mixture of low and high risk populations. Therefore, as a more flexible model for the true baseline risks, we will also consider in this chapter a mixture of two normal distributions with the same variance:

$$\xi_i \cong p_1 \mathrm{N}(\bar{\xi}_1, \sigma_\xi^2) + (1-p_1)\mathrm{N}(\bar{\xi}_2, \sigma_\xi^2) \qquad (3)$$

This model can describe a very broad class of distributions: unimodal as well as bimodal, symmetric as well as very skewed[27].

Thompson et al.[17] did not explicitly specify a parametric model for G. They specified independent vague mean normal prior distributions for the ξ_i's, which means that G is assumed to be N(ξ, σ^2) for some specified value ξ, e.g. ξ=0, and some large specified value of σ^2, e.g. σ^2=100. In a recent letter to the editor van Houwelingen en Senn[28] show that this method does not remove the bias that is present in the WLS approach. The intuitive argument is as follows. The slope of the regression line is equal to covar(ξ_i, η_i)/var(ξ_i). The WLS method is biased because the denominator is overestimated, since, due to sampling variability, the $\hat{\xi}_i$'s are more variable than the true ξ_i's. Therefore, to estimate the variance of the true ξ_i's, the estimates of them should be shrunk to the mean. However this is not accomplished by assuming that the ξ_i's are drawn from a normal distribution with a very large variance. On the contrary, it can lead to an estimate of var(ξ_i) is larger than the WLS estimate, and consequently to a larger bias in the slope.

The third model component models the within trial sampling variability in the estimates of ξ_i and η_i. McIntosh[4] and van Houwelingen et al.[20] assumed the following approximate normal model for the measurement errors:

$$\begin{pmatrix} \hat{\xi}_i \\ \hat{\eta}_i \end{pmatrix} \cong \mathrm{N}\left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \hat{\Sigma}\right) \qquad (4)$$

where $\hat{\Sigma}$ is the estimated covariance matrix of $\hat{\xi}_i$ and $\hat{\eta}_i$. The method is approximate in the sense that $\hat{\Sigma}$ is assumed to be known, i.e. no allowance is made for the fact that it is estimated. In this chapter we will use the exact distribution of $(\hat{\xi}_i, \hat{\eta}_i)$ given (ξ_i, η_i). In the case that the simple normal model (2) is assumed for the distributions of the true baseline risks, we will also fit the model with the above approximate measurement error model and compare the results. In the remainder of this section we describe in detail the models used in our three examples.

3.1 Model for example 1

The numbers of pre-term deliveries in the placebo and active treatment group of the i^{th} trial are denoted by X_i and Y_i. The corresponding sample sizes are m_i and n_i. Let ξ_i and η_i stand for the true placebo log odds and the true log odds ratio, respectively. The model for the relation between the true log odds ratio and the logit of the true baseline risk is given by (1). For the distribution of the logit of the true baseline risks

we will consider the simple normal model (2) and the mixture of two normal distributions (3).

The exact measurement model is implicitly given by assuming that X_i and Y_i have a binomial distribution:

$$X_i \cong \text{Binomial}(m_i, \frac{\exp(\xi_i)}{1+\exp(\xi_i)})$$

$$Y_i \cong \text{Binomial}(n_i, \frac{\xi_i \exp(\eta_i)}{1-\xi_i+\xi_i \exp(\eta_i)})$$

The approximate measurement model is:

$$\begin{pmatrix} \hat{\xi}_i \\ \hat{\eta}_i \end{pmatrix} \cong \text{N}\left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \begin{bmatrix} \frac{1}{x_i}+\frac{1}{m_i-x_i} & -\frac{1}{x_i}-\frac{1}{m_i-x_i} \\ -\frac{1}{x_i}-\frac{1}{m_i-x_i} & \frac{1}{x_i}+\frac{1}{m_i-x_i}+\frac{1}{y_i}+\frac{1}{n_i-y_i} \end{bmatrix}\right)$$

where $\hat{\xi}_i = \log(X_i/(m_i - X_i))$ and $\hat{\eta}_i = \log(Y_i/(n_i - Y_i))-\log(X_i/(m_i - X_i))$ are the estimates of the true placebo log odds ξ_i and the true odds ratio η_i. Notice that the left upper corner of the covariance matrix is the square of the usual standard error of a log odds of a proportion and the right lower corner is Woolf's squared standard error for a log odds ratio (½ is added to all denominators when one of them is zero). The covariance was computed using the usual approximate methods.

3.2 Model for example 2

The numbers of events in the placebo and treated group of the i^{th} trial are denoted by X_i and Y_i, respectively. The corresponding numbers of person years are m_i and n_i. Let ξ_i and η_i stand for the true log event rate under placebo and treatment, respectively. The model for the relation between the true log event rates is given by (1). For the distribution of the true log baseline rates we will consider the simple normal model (2) and the mixture of two normal distributions (3).

The exact measurement model is implicitly given by:

$$X_i \cong \text{Poisson}(m_i \cdot \exp(\xi_i))$$

$$Y_i \cong \text{Poisson}(n_i \cdot \exp(\eta_i))$$

Since the standard error of an estimated log event rate is equal to the inverse of the square root of the number of observed events, the approximate measurement model is:

$$\begin{pmatrix} \hat{\xi}_i \\ \hat{\eta}_i \end{pmatrix} \cong \mathrm{N}\left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \begin{pmatrix} \frac{1}{x_i} & 0 \\ 0 & \frac{1}{y_i} \end{pmatrix}\right)$$

3.3 Model for example 3

In the original analysis of Davey Smith et al.[7] the dependent variable in the regression was the log odds ratio for total mortality. Since the mean lengths of follow-up varied substantially among studies, we prefer the log of the event rate ratio, where the observed event rates in both groups are estimated as the number of events divided by the number of person years of follow-up. As independent variable we use the CHD mortality event rate under the control treatment.
The numbers of death from coronary heart disease in the placebo and active treatment group of the i^{th} trial are denoted by X_i and Y_i. The corresponding numbers of person years are m_i and n_i. Let ξ_i and η_i stand for the true placebo log events rate and the true log rate ratio, respectively. The model for the relation between the true log rate ratio and the baseline log event rate is again given by (1). For the distribution of the true baseline log event rate we will consider the simple normal model (2) and the mixture of two normal distributions (3).

The exact measurement model is implicitly given by:
$X_i \cong \mathrm{Poisson}(m_i \cdot \exp(\xi_i))$
$Y_i \cong \mathrm{Poisson}(n_i \cdot \exp(\eta_i + \xi_i))$

The approximate measurement model is:

$$\begin{pmatrix} \hat{\xi}_i \\ \hat{\eta}_i \end{pmatrix} \cong \mathrm{N}\left(\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix}, \begin{bmatrix} \frac{1}{x_i} & -\frac{1}{x_i} \\ -\frac{1}{x_i} & \frac{1}{x_i} + \frac{1}{y_i} \end{bmatrix}\right)$$

4 Method of estimation

When an exact measurement model is assumed, the model is difficult to fit using standard likelihood methods. Certainly no standard software can be used. Following Thompson et al.[17], we therefore adopted a Bayesian approach, using the BUGS implementation of Markov Chain Monte Carlo (MCMC) numerical techniques (Gibbs sampling)[21]. It turns out that it is relatively simple to carry out a fully Bayesian analysis with this BUGS program. A fully Bayesian analysis places prior distributions on all unknown parameters in the model. We used priors that are non-informative over the region supported by the likelihood. The BUGS code needed for the three examples above is given and annotated in Appendix 1. In a Bayesian analysis using MCMC methods it is relatively straightforward to get the posterior distribution of derived parameters. We used this to obtain confidence bands for the regression line and the break-even point (i.e. the baseline risk corresponding with no treatment effect). See Appendix 1 for how this was done using BUGS. Note that the confidence and prediction bands are evaluated at only ten points, which will usually be sufficient to draw smooth bands in a figure. Of course, more points might be chosen if considered desirable.

McIntosh[4] and van Houwelingen et al.[20] assumed the approximate measurement model together with the simple normal model for the baseline risk measure. In that case, straightforward (approximate) likelihood methods can be applied. McIntosh[4] and van Houwelingen et al.[20] both describe EM based algorithms, which unfortunately cannot be carried out using standard procedures from standard statistical packages. However, in Appendix 2 we show that it is relatively simple to fit this model using standard General Linear Mixed Model programs, provided the program has the option to keep certain covariance parameters fixed. We used the procedure Proc Mixed of SAS[29]. The code that is needed is given and explained in Appendix 2. Proc Mixed does not give a direct estimate of the slope and intercept of the regression line, but these are easily computed from the estimated covariance matrix of (ξ,η) given in the output. The slope is estimated as $\hat{\text{cov}}\text{ar}(\xi,\eta)/\hat{\text{var}}(\xi)$. The corresponding standard error is computed using the estimated covariance matrix of the estimated covariance matrix of (ξ,η) together with the delta method. The intercept and its standard error are computed analogously.

5 Results

On each of the three meta-analyses examples introduced in section 2 we applied five statistical models to examine the relationship between treatment effect and baseline risk. The results for each of the models are presented in Table 4. The first model (I) is the conventional WLS regression approach. Secondly, the results for the full Bayesian model (II) with the fixed, flat normal prior distribution on the baseline risks are given (method of Thompson et al.[17]). Subsequently, the results of our first Bayesian model (III), with a normal distribution for the baseline risks (2), are presented. Next, the results of our second model (IV) are given, with a mixture of two normal distributions as distribution for the baseline risks (3). Finally, the results of the approximate likelihood method of McIntosh[4] or van Houwelingen et al.[20] (model V) are presented with the aim to compare the results of the approximate likelihood method with our Bayesian approach using an exact measurement error model.

In the first example the log odds ratio is regressed on the baseline log odds. Note that the slope of the WLS line we have calculated differs from the one presented by Brand and Kragt[1], because we use the log odds of the baseline risk instead of the baseline risk itself. The same applies to the second and third examples, where we have log-transformed the rates in both the treatment and control groups. In the second example we relate the log mortality rate in the treatment group to the log mortality rate in the control group, like in the l'Abbé-plot in the meta-analysis of Hoes et al.[6]. So, in this second example, the slope of the regression line is tested against one (e.g. slope of the no-effect line) instead of zero. Finally, in third example we relate the effect size, measured as the log rate ratio, to the log of the death rate (per 1000 person years) in the control group.

The most important results here concern the slope of the regression line of effect size on baseline risk with its standard error and its confidence interval. In all three examples the slope of the WLS regression line is significantly negative, but we know that these slopes are biased. In all three examples this bias is not removed by model II, the Bayesian method with the flat prior distribution on the true baseline risks. The estimated slopes are even slightly more biased than the WLS estimates. This is reflected by the fact that the estimated standard deviations of the true baseline risks for this model are larger than the standard deviations of the observed baseline risks that are implicitly used in the WLS approach. In the first example, the standard deviation of the baseline risks increase from 1.05 in the WLS approach to 1.08 in the Bayesian model with the flat prior on the true baseline risks. In the second example the standard deviation increases from 0.84 to 1.13 and in the third example the standard deviation of the baseline risks increases from 1.11 to 1.22. As expected, the

Table 4. Results of analyses for the three meta-analysis examples.
Meta-analyses Brand & Kragt[1] : Relationship between treatment effect (log odds ratio) and log odds of baseline risk
Meta-analysis Hoes et al.[6] : Relationship between log event rate in treatment group versus log baseline rate
Meta-analyses Davey Smith et al.[7] : Relationship between treatment effect (log rate ratio) and log baseline rate

Meta-analysis	Model	Slope (se)	95% (Bayesian) confidence interval of slope	Intercept (95% (Bayesian) confidence interval)	Residual heterogeneity (τ)
Brand & Kragt	I WLS Regression	-0.44 (0.19)	(- 0.86 to -0.01)	-1.37 (-1.77 to -0.97)	0.86
13 trials					
(H_0 :slope = 0)	Model for baseline log odds:				
	II Fixed, flat prior	-0.58 (0.28)	(- 1.20 to -0.05)	-1.66 (-2.49 to -1.13)	0.47
	III Normal distribution	-0.25 (0.44)	(- 1.12 to 0.60)	-1.47 (-2.28 to -0.86)	0.37
	IV Mixture of two normals	-0.30 (0.46)	(- 1.14 to 0.77)	-1.50 (-2.33 to -0.83)	0.39
	V Approximate likelihood	-0.26 (0.33)	(- 0.90 to 0.38)	-1.35 (- 1.99 to -0.71)	0.13
Hoes et al.	I WLS Regression	0.64 (0.12)	(0.38 to 0.90)	0.65 (0.11 to 1.19)	1.05
12 trials					
(H_0:slope = 1)	Model for baseline log rate:				
	II Fixed, flat prior	0.58 (0.14)	(0.29 to 0.84)	0.76 (0.20 to 1.37)	0.15
	III Normal distribution	0.68 (0.13)	(0.42 to 0.95)	0.54 (- 0.03 to 1.08)	0.14
	IV Mixture of two normals	0.69 (0.14)	(0.42 to 0.97)	0.52 (- 0.06 to 1.09)	0.14
	V Approximate likelihood	0.69 (0.11)	(0.48 to 0.90)	0.53 (- 0.05 to 1.12)	0.10
Davey Smith et al.	I WLS Regression	-0.11 (0.04)	(-0.19 to -0.024)	0.36 (0.07 to 0.65)	1.12
33 trials					
(Total mortality)	Model for baseline log rate:				
(H_0: slope = 0)	II Fixed, flat prior	-0.11 (0.05)	(-0.21 to -0.021)	0.34 (0.01 to 0.69)	0.15
	III Normal distribution	-0.08 (0.05)	(-0.18 to 0.008)	0.24 (- 0.11 to 0.58)	0.14
	IV Mixture of two normals	-0.08 (0.05)	(-0.18 to 0.008)	0.23 (- 0.10 to 0.58)	0.13
	V Approximate likelihood	-0.08 (0.04)	(-0.17 to 0.002)	0.23 (- 0.31 to 0.77)	0.14

estimated slopes of the other three methods are shrunk towards zero (for example 1 and 3) or towards 1 (example 2) compared to the WLS approach. Particularly in the Brand & Kragt example the amount of shrinkage is large due to some small trials with relatively large within study variance compared to the between study variance (standard deviation of the observed baseline risks is 1.05, compared to 0.70 for the true baseline risks), resulting in a non-significant slope for all three methods. The shrinkage of the slope is very modest in the Hoes example (standard deviation of the observed baseline risks is 0.84 compared to 0.71 for the true baseline risks), due to the fact that the within trial variance of the baseline risks is relatively small compared to variation between trials, and the estimated slope remains statistically significant for all three methods. In the Davey Smith example, the shrinkage causes a reduction in the slope of about 25% for all three methods, and the slope is no longer statistically significant.

The difference between modelling the baseline risk distribution with one normal distribution (model III) or a mixture of two normal distributions (model IV) turns out to be negligible in all three examples. The estimated regression lines of the approximate likelihood approach (model V) turn out to agree very well with the two Bayesian models III and IV. The standard errors of the slope of the approximate likelihood approach are somewhat smaller, probably reflecting the fact that this method does not account for the uncertainty in the standard errors of the estimates of treatment effects and baseline risks.

For the three Bayesian methods the estimated break-even points for the Hoes example are presented in Table 5. Since the posterior distributions were rather skewed, the break-even points are estimated by the median of their posterior distribution. Note that the break-even point in model II (fixed flat prior) is very similar to the 6 / 1000 pryrs calculated by Hoes et al. with a WLS regression analysis. However, again the results of model II are biased because of the bias of the corresponding estimated regression line. This point estimate becomes little lower when estimated by model III and IV, suggesting a somewhat larger group of patients for which treatment has a beneficial effect. Only minor differences exist between the estimates of the break-even points and the corresponding confidence intervals of models III and IV. Notice that the 95% confidence interval for models III and IV are somewhat larger than for model II, indicating a larger range of baseline risks for which treatment effect is uncertain. In this example of Hoes et al.[6], the 95% confidence interval stays strictly positive when analysed with model III and IV, implying that estimated break-even point is significantly positive. This suggests that there indeed exists a baseline event rate below which treatment may have harmful effects. As an illustration of model III, we

plotted in Figure 4 the estimated regression line together with its 95% Bayesian confidence band and the break-even point for the meta-analysis of Hoes et al.[6].

Table 5. Point of intersection (break-even point) of the no-effect line with the regression line where the log event rate in the treatment group is plotted against the log event rate in the control group (baseline risk).

Meta-analysis	Model	Median log break-even point ξ	95% interval log break-even point ξ	Median baseline-risk or rate	95% interval baseline-risk or rate
Hoes et al.	Model for baseline log rate:				
12 trials	II Fixed, flat	1.80	(1.15 to 2.17)	6.05 / 1000 prys	(3.16 to 8.76 / 1000 prys)
	III Normal	1.73	(0.41 to 2.25)	5.64 / 1000 prys	(1.51 to 9.49 / 1000 prys)
	IV Mixture two normals	1.71	(0.12 to 2.29)	5.53 / 1000 prys	(1.13 to 9.87 / 1000 prys)

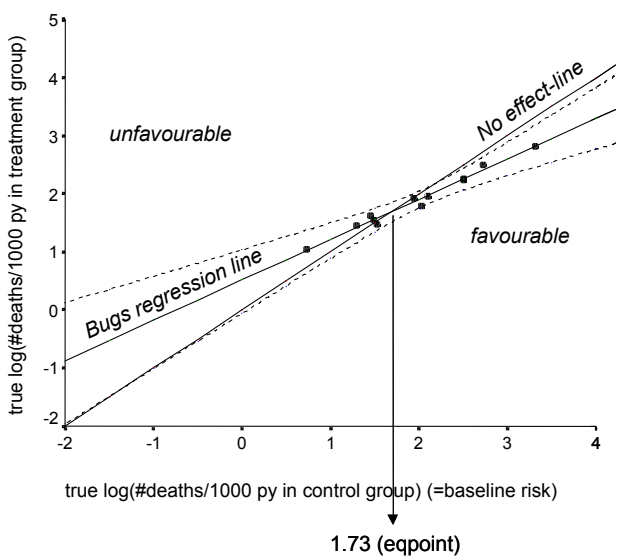


Figure 4. Meta-analysis Hoes et al.[6] with the Bugs (true) regression line (model III with normal distribution on baseline risks) and the 95% Bayesian confidence interval around the regression line.

6 Discussion

In this chapter we have proposed a hierarchical Bayesian modelling approach to investigate the relation between treatment effect and baseline risk in clinical trial meta-analyses. The three meta-analysis examples show that the results can differ substantially from the standard weighted linear regression approach. Our method can be considered as an generalisation of the method proposed by Thompson et al.[17]. In agreement with the theoretical argument of van Houwelingen and Senn[28] our results show that more extended modelling of the distribution of the baseline risks is preferable.

Our results were very comparable to the results of McIntosh's method, although this method gave somewhat smaller standard errors, probably due to the fact that no allowance is made for the uncertainty in the trial specific standard errors. An advantage of the method of McIntosh is that it can be carried out in at least one widely available standard statistical package, although for the standard errors of slope and intercept of the regression line some extra programming is needed.

The method we propose in this chapter has several advantages. It can be carried out using the freely available Bayesian statistical package BUGS[21], which enables very flexible modelling. An exact measurement error model can be specified. Sensitivity analyses on the assumptions of the model are easily performed. For instance, to investigate whether a normal distribution for the baseline risks is a plausible assumption, we also fitted a model that assumed a mixture of two normal distributions. Many other distributional assumptions could be tried as well. Other model assumptions could be varied too. For instance, instead of the normal distribution for the residuals around the regression line, a *t*-distribution could be specified. Or, instead of a linear relation, a quadratic relationship could be assumed. Another advantage of our method is that for derived parameters almost automatically confidence intervals are obtained, such as a confidence interval for the break-even point or a confidence band for the underlying relationship between the true effect measure and baseline risk. Also a prediction interval, giving the range in which the treatment effects probably will lie for a new trial in a population with a given baseline risk, can be obtained. Additionally, the model can easily be expanded with more covariables. Also the grouped random effect models for analysis, as suggested by Larose & Day[30] to distinguish between distinct kinds of studies can be naturally implemented in our approach, as long as one has enough trials of reasonable size.

As a final remark, we do not fully agree with the point of view of van Houwelingen and Senn[28] that there is only limited direct relevance of our analyses to the decision making process for individual patients and their doctors. Of course, patients as well

as doctors need to know which measurable personal features are beforehand related to the success of treatment to make better decisions, instead of calculating the baseline risk of the patients in a certain clinical practice in retrospect as we did in this chapter. However, in our opinion there is still a place for the kind of analyses considered in this chapter. First, even without knowing the exact or approximate baseline risk of a specific patient, it is useful for a doctor to know whether the treatment works better as the risk of a patient is higher. If there is such a relationship, the doctor can pay attention to general or specific risk factors like age to advise treatment or not. When no relationship whatsoever between treatment effect and baseline risk can be demonstrated, the doctor can base his treatment advice on other reasons, not taking risk factors into account. However, note that when the relative risk is constant over baseline risk, the absolute risk-reduction can still be larger at increasing baseline risk. This implies that lower numbers to treat are needed to save one person or in other words, that the same effect could be achieved at lower costs. Therefore, in that case the doctor could still decide to treat only the high risk patients because of cost-effectiveness arguments, although the relative risk is independent of baseline risk. Second, again even without knowing the baseline risk of a specific patient, it can be important for a doctor to know whether there are patients who have a baseline risk below some break-even point for whom treatment has no influence or may even have a harmful effect. When the estimated break-even point according to our model(s) appears to be at an extremely low risk of which we know that actually all people fall above this point, doctors need not be afraid for harmful effects for the patient. At the same time, for a treatment with a very high break-even point, doctors need to be careful in prescribing treatment, especially when the endpoint is a serious one. In the last case the following strategy, suggested by Sharp et al.[16] and by Thompson et al.[17], should certainly be considered. They suggest a two-step approach: 1. executing an analysis like the one presented in this chapter of the relationship between treatment effect and baseline risk based on trial-level data and 2. developing a prognostic model based on a large pool individual follow-up data to combine several prognostic variables to predict the 'baseline' risk. With these two steps, one can approximately predict risk and hence treatment benefit for an individual patient.

Appendix 1

The BUGS code <- means 'is equal to' and ~ means 'is distributed as'. From the model specification BUGS constructs the necessary full conditional distributions and carries out Gibbs sampling. For the examples presented here a 'burn-in' of 5.000 iterations was followed by further 15.000 iterations during which the generated parameter values were monitored and summary statistics as the median and 95% credible interval of the complete samples were obtained. Note that BUGS parameterises the normal distribution (dnorm) in terms of the precision rather than the variance.

BUGS code (EXAMPLE 1 and 3):

```
# Step 1: Underlying regression model:
for(i in 1:13 ){mean.eta[i] <- alpha + (beta+1)*(ksi[i]-mean(ksi[]));}
for(i in 1:13 ){eta[i] ~ dnorm(mean.eta[i],tau);}
 # with the vague priors:
tau ~ dgamma(0.001,0.001);
beta ~ dnorm(0.0,1.0E-6);
alpha ~ dnorm(0.0,1.0E-6);

# Step 2: Baseline risks model:
# Model 1 (vague prior): for( i in 1 : 13 ) {ksi[i] ~ dnorm(0.0,0.001);}

# Model 2 (normal empirical Bayes prior):
for( i in 1 : 13 ) {ksi[i] ~ dnorm(mean.ksi,tau.ksi);}
  # with the vague priors:
mean.ksi ~ dnorm(0, 0.001) and  tau.ksi ~ dgamma(0.001,0.001);

# Model 3 (mixed empirical Bayes prior):
for( i in 1 : 13 ) {  ksi[i] ~ dnorm(mean.ksi[T[i]],tau.ksi);}
for( i in 1 : 13 ) {  T[i] ~ dcat(P[]) }
   # with vague priors:
tau.ksi  ~ dgamma(0.001,0.001);
mean.ksi[1] ~ dnorm(0,1.0E-6);
mean.ksi[2] <- mean.ksi[1] + theta;
theta ~ dnorm(0,1.0E-6)I(0,);
P[]~ddirch(alfa[]); alfa[1]<-1; alfa[2]<-1;

# Step 3: Measurement errors model:
for( i in 1 : 13 ) {y[i] ~ dbin(lambda[i],n[i]);}
for( i in 1 : 13 ) {x[i] ~ dbin(mu[i],m[i]);}
```

 # and transforming the risks into logit(risks) for control and treatment group:

```
for( i in 1 : 13 ) {logit(mu[i]) <- ksi[i];}
for( i in 1 : 13 ) {logit(lambda[i]) <- eta[i];}
```

 # Easy to calculate the value of the intercept of the non-centred regression line and the value # of the baseline risk measure for which there is no treatment effect (eqpoint):

```
alpha.real <- alpha-(beta+1)* mean(ksi[]);
eqpoint <- (alpha.real)/(-beta);
```

To calculate a confidence and a prediction band for the expected treatment effect given a true baseline risk:

```
for (i in 1:10) {m.eta[i] <- alpha + (beta+1)*(i*0.5-mean(ksi[]));}
for (i in 1:10) {new[i] ~ dnorm(m.eta[i] , tau);}
```

BUGS code (EXAMPLE 2):

Underlying regression model:

```
for( i in 1 : 12 ) {mean.eta[i] <- alpha + beta * (ksi[i]-mean(ksi[]));}
for( i in 1 : 12 ) {eta[i] ~ dnorm(mean.eta[i],tau);}
```

with the vague priors:

```
tau ~ dgamma(0.001,0.001);
beta ~ dnorm(0.0,1.0E-6);
alpha ~ dnorm(0.0,1.0E-6);
```

Baseline risks model, model 1 (vague prior):

```
for( i in 1 : 12 ) {ksi[i] ~ dnorm(0.0,0.001);}
```

Baseline risks model, model 2 (normal empirical Bayes prior):

```
for( i in 1 : 12 ) {ksi[i] ~ dnorm(mean.ksi,tau.ksi);}
```

with the vague priors:

```
for( i in 1 : 12 ) {ksi[i] ~ dnorm(0.0,0.00001); }
tau.ksi ~ dgamma(0.001,0.001);
```

Baseline risks model, model 3 (mixed empirical Bayes prior):

```
for( i in 1 : 12 ) {  ksi[i] ~ dnorm(mean.ksi[T[i]],tau.ksi);}
for( i in 1 : 12 ) {  T[i] ~ dcat(P[]) }
```

with vague priors:

```
tau.ksi  ~ dgamma(0.001,0.001);
mean.ksi[1] ~ dnorm(0,1.0E-6);
mean.ksi[2] <- mean.ksi[1] + theta;
theta ~ dnorm(0,1.0E-6) I(0,);
P[] ~ ddirch(alfa[]);
alfa[1] <- 1;
alfa[2] <- 1;
```

Measurement errors model:

```
for( i in 1 : 12 )  {y[i] ~ dpois(lambda[i]);}
for( i in 1 : 12 )  {x[i] ~ dpois(mu[i]); }
```

and transforming the risks into log(rates) for control and treatment group:

```
for( i in 1 : 12 ) {log(mu[i]) <- log(n[i]) + ksi[i];}
for( i in 1 : 12 ) {log(lambda[i]) <- log(m[i]) + eta[i];}
```

Easy to calculate the value of the intercept of the non-centred regression line and the value of the baseline risk measure for which there is no treatment effect:

```
alpha.real<-alpha-beta*mean(ksi[]);
eqpoint <- alpha.real/(1-beta);
```

And to calculate a confidence and a prediction band for the expected treatment effect given a true baseline risk:

```
for (i in 1:10) {m.eta[i] <- alpha + beta*(i*0.5-mean(ksi[]));}
for (i in 1:10) {new[i] ~ dnorm(m.eta[i] , tau);}
```

Appendix 2

Below the SAS Proc Mixed code as used for example 2 is given. The data set 'hoes' contains a data record for each treatment arm for all trials. The variables are:

trial = *trial number, same within one trial (in this example 1 to 12),*

grp = *successive number of treatment arm, different between control and experimental group within one trial and between trials (here: 1 to 24),*

logodds = *estimated log odds per trial-arm,*

exp = *1 for experimental group, 0 for control group,*

con = *1 for control group, 0 for experimental group*

The SAS commands are:

```
proc mixed method=ml cl data=hoes;           * Call procedure.
  class trial grp;                           * Specifies study and
                                             successive number of
                                             treatment-arm as
                                             classification var.
  model logor= exp con / noint s cl;         * Model with indication
                                             variables 'exp' and 'con' as
                                             explanatory variables for
                                             logodds. Print solution 's'.
  random exp con / subject=trial type=un s;  * Trial is specified as
                                             random effect;
                                             Covariance matrix is
                                             unstructured; Print
                                             empirical Bayes estimates
                                             's'.
  repeated / group=grp;                      * Each study-arm in each
                                             trial has its own within
                                             study error variance
  parms / parmsdata=covvars eqcons=4 to 27;  * Starting values of
                                             covariance
                                             parameters: the two between
                                             study variances, its
                                             covariance and the 26 within
                                             study error variances;
                                             Considering the last 26
                                             values known and constant.
run;
```

The SAS-datafile 'covvars' has one variable, of which the first three values are the starting values for

a. the between study variance of the log odds in the experimental group,

b. the covariance between the true log odds in control and experimental group,

c. the between study variance of the log odds of the control group.

The next 24 values are the estimated variances of the log odds within alternately the experimental and the control groups, computed as $(\frac{1}{y_i}+\frac{1}{n_i-y_i})$ and $(\frac{1}{x_i}+\frac{1}{m_i-x_i})$, respectively, where x_i and y_i are the number of events in the control respectively the active treatment group of the i^{th} trial with corresponding sample sizes m_i and n_i. The subcommand 'eqcons=4 to 27' means that these 24 variances are kept fixed, i.e. treated as known.

References

1. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* 1992; **11**(16):2077-2082.
2. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**(6965):1351-1355.
3. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177-188.
4. McIntosh MW. The population risk as an explanatory variable in research syntheses of clinical trials. *Statistics in Medicine* 1996; **15**:1713-1728.
5. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**:2685-2699.
6. Hoes AW, Grobbee DE, Lubsen J. Does drug treatment improve survival? Reconciling the trials in mild-to-moderate hypertension. *Journal of Hypertension* 1995; **13**(7):805-811.
7. Davey Smith G, Song F, Sheldon TA. Cholesterol lowering and mortality: the importance of considering initial level of risk. *British Medical Journal* 1993; **306**(6889):1367-1373.
8. Berlin JA, Antman EM. Advantages and limitations of meta-analytic regressions of clinical trials data. *Online Journal of Current Clinical Trials* 1994; **134**.
9. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet* 1995; **345**(8965):1616-1619.
10. Boissel J, Collet J, Lievre M, Girard P. An effect model for the assessment of drug benefit: example of antiarrhythmic drugs in postmyocardial infarction patients. *Journal of Cardiovascular Pharmacology* 1993; **22**:356-363.
11. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Journal of American Medical Association* 1992; **268**:240-248.
12. Lau J, Chalmers TC. Should magnesium be given for acute myocardial infarction? *Clinical Research* 1994; **42**:290A.
13. Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). *Statistics in Medicine* 1994; **13**(3):293-295.
14. Egger M, Smith GD. Risks and benefits of treating mild hypertension: a misleading meta-analysis? [comment]. *Journal of Hypertension* 1995; **13**(7):813-815.
15. Cook RJ, Walter SD. A logistic model for trend in 2 x 2 x K tables with applications to meta-analyses. *Biometrics* 1997; **53**(1):352-357.

16. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735-738.
17. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741-2758.
18. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2883-2900.
19. Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analysis. *Statistics in Medicine* 1994; **13**:1265-1282.
20. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**(24):2273-2284.
21. Spiegelhalter D, Thomas A, Best N, Gilks W. *BUGS: Bayesian inference Using Gibbs Sampling (version 0.6)*. MRC Biostatistics Unit, Institute of Public Health: Cambridge, 1996.
22. Keirse MJNC, Grant A, King JF. Preterm labour. *Effective care in pregnancy and childbirth.* In Chalmers I, Enkin M, Keirse MJNC (ed.). Oxford University Press: Oxford, 1989; pp 694-745.
23. Hoes AW, Grobbee DE, Stijnen T, Lubsen J. Meta-analysis and the Hippocratic principle of primum non nocere [Authors' reply]. *Journal of Hypertension* 1995; **13**:1353-1355.
24. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Annals of Internal Medicine* 1987; **107**(2):224-233.
25. MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, et al. Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990; **335**(8692):765-774.
26. Bernsen RMD, Tasche MJA, Nagelkerke NJD. Some notes on baseline risk and heterogeneity in meta-analysis. *Statistics in Medicine* 1999; **18**(2):233-238.
27. Verbeke G. Linear Mixed Models for Longitudinal Data. *Linear Mixed Models in Practice*. In: Verbeke G, Molenberghs G (ed.). Springer-Verlag: New York, 1997; pp 63-153.
28. Van Houwelingen HC, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis (letter). *Statistics in Medicine* 1999; **18**:107-113.
29. SAS Institute Inc. *SAS/STAT User's guide, Version 6.* Cary NC, 1989.
30. Larose DT, Dey DK. Grouped random effects models for Bayesian meta-analysis. *Statistics in Medicine* 1997; **16**(16):1817-1829.

3

Advanced methods in meta-analysis: multivariate approach and meta-regression

Abstract

This tutorial on advanced statistical methods for meta-analysis can be seen as a sequel to the recent Tutorial in Biostatistics on meta-analysis by Normand[1], which focussed on elementary methods. Within the framework of the general linear mixed model using approximate likelihood, we discuss methods to analyse univariate as well as bivariate treatment effects in meta-analyses as well as meta-regression methods. Several extensions of the models are discussed, like exact likelihood, non-normal mixtures and multiple endpoints. We end with a discussion about the use of Bayesian methods in meta-analysis. All methods are illustrated by a meta-analysis concerning the efficacy of BCG vaccine against tuberculosis. All analyses that use approximate likelihood can be carried out by standard software. We demonstrate how the models can be fitted using SAS Proc Mixed.

1 Introduction

In this chapter we review advanced statistical methods for meta-analysis as used in bivariate meta-analysis[2] (i.e. two outcomes per study are modelled simultaneously) and meta-regression[3]. It can be seen as a sequel to the recent Tutorial in Biostatistics on meta-analysis by Normand[1]. Meta-analysis is put in the context of mixed models using (approximate) likelihood methods to estimate all relevant parameters. In the medical literature meta-analysis is usually applied to the results of clinical trials, but the application of the theory presented in this chapter is not limited to clinical trials only. It is the aim of this chapter not only to discuss the underlying theory but also to give practical guidelines how to carry out these analyses.

As leading example we use the meta-analysis data set of Colditz et al.[4]. This data set is also discussed in Berkey et al.[3]. Wherever feasible, it is specified how the analysis can be performed by using the SAS procedure Proc Mixed. The chapter is organised as follows. In section 2 we review the concept of approximate likelihood that was introduced in the meta-analysis setting by DerSimonian & Laird[5]. In section 3 we review the meta-analysis of one-dimensional treatment effect parameters. In section 4 we discuss the bivariate approach[2] and its link with the concept of underlying risk as source of heterogeneity[6-10]. In section 5 we discuss meta-regression within the mixed model setting. Covariates considered are aggregate measures on the study level. We do not go into meta-analysis with patient-specific covariates. In principle that is not different from analysing a multi-centre study[11]. In section 6 several extensions are discussed: exact likelihood's based on conditioning, non-normal mixtures, multiple endpoints, other outcome measures and other software. This is additional material that can be skipped at first reading. Section 7 is concerned with the use of Bayesian methods in meta-analysis. We argue that Bayesian methods can be useful if they are applied at the right level of the hierarchical model. The chapter is concluded in section 8.

2 Approximate Likelihood

The basic situation in meta-analysis is that we are dealing with n studies in which a parameter of interest ϑ_i (i=1,..., n) is estimated. In a meta-analysis of clinical trials the parameter of interest is some measure of the difference in efficacy between the two treatment arms. The most popular choice is the log odds ratio, but this could also be the risk- or rate-difference or the risk- or rate-ratio for dichotomous outcome or similar measures for continuous outcomes or survival data. All studies report an

estimate $\hat{\vartheta}_i$ of the true ϑ_i and the standard error s_i of the estimate. If the studies only report the estimate and the *p*-value or a confidence interval, we can derive the standard error from the *p*-value or the confidence interval. In the sections 3 to 5, which give the main statistical tools, we act as if $\hat{\vartheta}_i$ has a normal distribution with unknown mean ϑ_i and known standard deviation s_i, that is

$$\hat{\vartheta}_i \sim \mathrm{N}(\vartheta_i, s_i^2) \qquad (1)$$

Moreover, since the estimates are derived from different data sets, the $\hat{\vartheta}_i$ are conditionally independent given ϑ_i. This approximate likelihood approach goes back to the seminal paper by DerSimonian & Laird[5]. However, it should be stressed that it is <u>not</u> the normality of the frequency distribution of $\hat{\vartheta}_i$ that is employed in our analysis. Since our whole approach is likelihood based, we only use that the likelihood of the unknown parameter in each study looks like the likelihood of (1). So, if we denote the log-likelihood of the *i*-th study by $\ell_i(\vartheta)$, the real approximation is

$$\ell_i(\vartheta) = -\frac{1}{2}(\vartheta - \hat{\vartheta}_i)^2/s_i^2 + c_i \qquad (2)$$

where c_i is some constant that does not depend on the unknown parameter.

If in each study the unknown parameter is estimated by Maximum Likelihood, approximation (2) is just the second order Taylor expansion of the (profile) log-likelihood around the MLE $\hat{\vartheta}_i$. The approximation (2) is usually quite good, even if the estimator $\hat{\vartheta}_i$ is discrete. Since most studies indeed use the Maximum Likelihood method to estimate the unknown parameter, we are confident that (2) can be used as an approximation. In section 6 we will discuss some refinements of this approximation. In manipulating the likelihood's we can safely act as if we assume that (1) is valid and use, for example, known results for mixtures of normal distributions. However, we want to stress that actually we only use assumption (2).

The approach of Yusuf et al.[12], popular in fixed effect meta-analysis, and of Whitehead and Whitehead[13] are based on a Taylor expansion of the log-likelihood around the value $\vartheta = 0$. This is valid if the effects in each study are relatively small. It gives an approximation in the line of (2) with different estimators and standard errors but a similar quadratic expression in the unknown parameter.

As we already noted, the most popular outcome measure in meta-analysis is the log odds ratio. Its estimated standard error is equal to ∞ if one of the frequencies in the 2×2 table is equal to zero. That is usually repaired by adding ½ to all cell frequencies. We will discuss more appropriate ways of handling this problem in section 6.

3 Analysing one-dimensional treatment effects

The analysis under 'homogeneity' makes the assumption that the unknown parameter is exactly the same in all studies, that is $\vartheta_1 = \vartheta_2 = ... = \vartheta_n = \vartheta$. The log-likelihood for ϑ is given by

$$\ell(\vartheta) = \sum_i \ell_i(\vartheta) = -\frac{1}{2}\sum_i [(\vartheta - \hat{\vartheta}_i)^2 / s_i^2 + \ln(s_i^2) + \ln(2\pi)] \qquad (3)$$

Maximisation is straight-forward and results in the well-known estimator of the common effect

$$\hat{\vartheta}_{hom} = [\sum_i \hat{\vartheta}_i / s_i^2] / [\sum_i 1 / s_i^2]$$

with standard error

$$se(\hat{\vartheta}_{hom}) = 1/\sqrt{\sum_i 1/s_i^2}$$

Confidence intervals for ϑ can be based on normal distributions, since the s_i^2 terms are assumed to be known. Assuming the s_i^2 terms to be known instead of to be estimated has little impact on the results[14]. This is the basis for the traditional meta-analysis.

The assumption of homogeneity is questionable even if it is hard to disprove for small meta-analyses[15]. That is, heterogeneity might be present and should be part of the analysis even if the test for heterogeneity is not significant. Heterogeneity is found in many meta-analyses and is likely to be present since the individual studies are never identical with respect to study populations and other factors that can cause differences between studies.

The popular model for the analysis under 'heterogeneity' is the normal mixture model, introduced by DerSimonian and Laird[5], that considers the ϑ_i to be an independent random sample from a normal population

$$\vartheta_i \sim \mathrm{N}(\vartheta, \sigma^2)$$

Normality of this mixture is a true assumption and not a simplifying approximation. We will further discuss it in section 6. The resulting marginal distribution of ϑ_i is easily obtained as $\hat{\vartheta}_i \sim N(\vartheta, \sigma^2 + s_i^2)$ with corresponding log-likelihood

$$\ell(\vartheta,\sigma^2) = -\frac{1}{2}\sum_i [(\vartheta - \hat{\vartheta}_i)^2/(\sigma^2 + s_i^2) + \ln(\sigma^2 + s_i^2) + \ln(2\pi)] \qquad (4)$$

Notice that (3) and (4) are identical if $\sigma^2 = 0$.

This log-likelihood is the basis for inference about <u>both</u> parameters ϑ and σ^2. Maximum Likelihood estimates can be obtained by different algorithms. In the example below, it is shown how the estimates can be obtained by using the SAS procedure Proc Mixed. If σ^2 were known, the ML estimate for ϑ would be

$$\hat{\vartheta}_{het} = [\sum_i (\hat{\vartheta}_i/(\sigma^2 + s_i^2)]/[\sum_i [1/(\sigma^2 + s_i^2)]$$

with standard error

$$se(\hat{\vartheta}_{het}) = 1/\sqrt{\sum_i 1/(\sigma^2 + s_i^2)}$$

The latter can also be used if σ^2 is estimated and the estimated value is plugged in, as is done in the standard DerSimonian and Laird approach.

The construction of confidence intervals for both parameters is more complicated than in the case of a simple sample from a normal distribution. Simple χ^2- and *t*-distributions with df=*n*-1 are not appropriate. In this chapter all models are fitted using SAS Proc Mixed, which gives Satterthwaite approximation based confidence intervals. Another possibility is to base confidence intervals on the likelihood ratio test, using profile log-likelihood's. That is, the confidence interval consists of all parameter values that are not rejected by the likelihood ratio test. Such confidence intervals often have amazingly accurate coverage probabilities[16, 17]. Brockwell and Gordon[18] compared the commonly used DerSimonian and Laird method[5] with the profile likelihood method. Particularly when the number of studies is modest, the DerSimonian and Laird method had coverage probabilities considerably below 0.95 and the profile likelihood method achieved the best coverage probabilities.

The profile log-likelihood's are defined by

$$p\ell_1(\vartheta) = \max_{\sigma^2} \ell(\vartheta,\sigma^2) \text{ and } p\ell_2(\sigma^2) = \max_{\vartheta} \ell(\vartheta,\sigma^2)$$

Based on the usual $\chi^2_{[1]}$-approximation for $2(p\ell_1(\hat{\vartheta}) - p\ell_1(\vartheta))$ the 95%- confidence interval for ϑ is obtained as all ϑ's satisfying $p\ell_1(\vartheta) > p\ell_1(\hat{\vartheta}) - 1.92$ (1.92 is the 95% centile of the $\chi^2_{[1]}$ distribution 3.84 divided by 2) and similarly for σ^2. Unlike the usual confidence interval based on Wald's method, this confidence interval for ϑ implicitly accounts for the fact that σ^2 is estimated.

Testing for heterogeneity is equivalent to testing $H_0 : \sigma^2 = 0$ against $H_1 : \sigma^2 > 0$. The likelihood ratio test statistic is $T = 2(p\ell_2(\hat{\sigma}^2) - p\ell_2(0))$. Since $\sigma^2 = 0$ is on the boundary of the parameter space, T does not have a $\chi^2_{[1]}$-distribution, but its distribution is a mixture with probabilities ½ of the degenerate distribution in zero and the $\chi^2_{[1]}$-distribution[19]. That means that the *p*-value of the naive LR-test has to be halved.

Once the mixed model has been fitted, the following information is available at the overall level:

- $\hat{\vartheta}$ and its confidence interval, showing the existence or absence of an overall effect
- $\hat{\sigma}^2$ and its confidence interval (and the test for heterogeneity), showing the variation between studies
- approximate 95% prediction interval for the *true* parameter $\hat{\vartheta}_{new}$ of a new unrelated study: $\hat{\vartheta} \pm 1.96\hat{\sigma}$. (Approximate in the sense that it ignores the error in the estimation of ϑ and σ)
- an estimate of the probability of a positive result of a new study: $P(\vartheta_{new} > 0) = \Phi(\hat{\vartheta}/\hat{\sigma})$
  (where Φ is the standard normal cumulative distribution function)

And at the individual level:

- posterior confidence intervals for the true ϑ_i's of the studies in the meta-analysis based on the posterior distribution $\vartheta_i \mid \hat{\vartheta}_i \sim \mathrm{N}(\hat{\vartheta} + B_i(\hat{\vartheta}_i - \hat{\vartheta}), B_i s_i^2)$ with $B_i = \hat{\sigma}^2/(\hat{\sigma}^2 + s_i^2)$. The posterior means or so-called empirical Bayes estimates give a more realistic view on the results of, especially, the small studies. See the meta-analysis tutorial of Normand[1] for more on this subject.

Example

To illustrate above methods we make use of the meta-analysis data given by Colditz et al.[4]. Berkey et al.[3] also used this dataset to illustrate their random-effects regression approach to meta-analysis. The meta-analysis concerns 13 trials on the efficacy of BCG vaccine against tuberculosis. In each trial a vaccinated group is compared with a non-vaccinated control group. The data consist of the sample size in each group and the number of cases of tuberculosis. Furthermore some covariates are available that might explain the heterogeneity among studies: geographic latitude of the place where the study was done, year of publication and method of treatment allocation (random, alternate, or systematic). The data are presented in Table 1.

Table 1. Example: Data from clinical trials on efficacy of BCG vaccine in the prevention of tuberculosis[3, 4].

	Vaccinated		Not vaccinated		ln(OR)	Latitude	Year	Allocation
Trial	Disease	No disease	Disease	No disease				
1	4	119	11	128	-0.93869	44	48	Random
2	6	300	29	274	-1.66619	55	49	Random
3	3	228	11	209	-1.38629	42	60	Random
4	62	13,536	248	12,619	-1.45644	52	77	Random
5	33	5,036	47	5,761	-0.21914	13	73	Alternate
6	180	1,361	372	1,079	-0.95812	44	53	Alternate
7	8	2,537	10	619	-1.63378	19	73	Random
8	505	87,886	499	87,892	0.01202	13	80	Random
9	29	7,470	45	7,232	-0.47175	27*	68	Random
10	17	1,699	65	1,600	-1.40121	42	61	Syst. Alloc.
11	186	50,448	141	27,197	-0.34085	18	74	Syst. Alloc.
12	5	2,493	3	2,338	0.44663	33	69	Syst. Alloc.
13	27	16,886	29	17,825	-0.01734	33	76	Syst. Alloc.

* This was actually a negative number, we used the absolute value in the analysis

We stored the data into a SAS-file called '`BCG_data.sd2`' (see Data step in SAS commands below). The treatment effect measure we have chosen is the log odds ratio, but the analysis could be carried out in the same way for any other treatment effect measure.

Fixed effects model

The analysis under the assumption of homogeneity is easily performed by hand. Only for the sake of continuity and uniformity we also show how the analysis can be carried out using SAS software.

The ML-estimate of the log odds ratio for trial *i* is:

$$\ln OR_i = \log\left(\frac{Y_{A,i}/(n_{A,i}-Y_{A,i})}{Y_{B,i}/(n_{B,i}-Y_{B,i})}\right)$$

where $Y_{A,i}$ and $Y_{B,i}$ are the number of disease cases in the vaccinated (A) and non-vaccinated group (B) in trial *i*, and $n_{A,i}$ and $n_{B,i}$ the sample sizes. The corresponding

within-trial variance, computed from the inverse of the matrix of second derivatives of the log-likelihood, is:

$\text{var}(\ln OR_i) = \frac{1}{Y_{\text{A},i}} + \frac{1}{n_{\text{A},i} - Y_{\text{A},i}} + \frac{1}{Y_{\text{B},i}} + \frac{1}{n_{\text{B},i} - Y_{\text{B},i}}$, which is also known as Woolf's formula.

These within-trial variances were stored in the same SAS data-file as above, called `'BCG_data.sd2'`. In the analysis, these variances are assumed to be known and fixed.

```
# THE DATA STEP;
data BCG_data;
input TRIAL VD VWD NVD NVWD LATITUDE YEAR ALLOC;
LN_OR=log((VD/VWD)/(NVD/NVWD));
EST=1/VD+1/VWD+1/NVD+1/NVWD;
datalines;
1    4       119        11       128         44         48         1
2    6       300        29       274         55         49         1
3    3       228        11       209         42         60         1
4   62     13536       248     12619         52         77         1
5   33      5036        47      5761         13         73         2
6  180      1361       372      1079         44         53         2
7    8      2537        10       619         19         73         1
8  505     87886       499     87892         13         80         1
9   29      7470        45      7232         27         68         1
10  17      1699        65      1600         42         61         3
11 186     50448       141     27197         18         74         3
12   5      2493         3      2338         33         69         3
13  27     16886        29     17825         33         76         3
;
proc print;run;
```

Running these SAS commands gives the following output:

```
OBS TRIAL VD   VWD   NVD   NVWD LATITUDE   YEAR  ALLOC    LN_OR      EST

 1    1    4    119   11    128      44      48     1    -0.93869    0.35712
 2    2    6    300   29    274      55      49     1    -1.66619    0.20813
 3    3    3    228   11    209      42      60     1    -1.38629    0.43341
 4    4   62  13536  248  12619      52      77     1    -1.45644    0.02031
 5    5   33   5036   47   5761      13      73     2    -0.21914    0.05195
 6    6   80   1361  372   1079      44      53     2    -0.95812    0.00991
 7    7    8   2537   10    619      19      73     1    -1.63378    0.22701
 8    8  505  87886  499  87892      13      80     1     0.01202    0.00401
 9    9   29   7470   45   7232      27      68     1    -0.47175    0.05698
10   10   17   1699   65   1600      42      61     3    -1.40121    0.07542
11   11  186  50448  141  27197      18      74     3    -0.34085    0.01253
12   12    5   2493    3   2338      33      69     3     0.44663    0.53416
13   13   27  16886   29  17825      33      76     3    -0.01734    0.07164
```

The list of variables matches that in Table 1 (*VD* = Vaccinated and Diseased, *VWD* = Vaccinated and Without Disease, *NVD* = Not Vaccinated and Diseased, *NVWD* = Not Vaccinated and Without Disease. The variable `ln_or` contains the estimated log odds

ratio of each trial and the variable `est` contains its variance per trial. In the Proc Mixed commands below, SAS assumes that the within trial variances are stored in a variable with the name '`est`'.

```
# THE FIXED EFFECTS MODEL;
Proc mixed method=ml                    #call SAS procedure;
data=BCG_data;
class trial;                            #specifies 'trial' as classification
                                        variable;
model ln_or=/ s ;                       #an intercept only model; print the
                                        solution s;
repeated /group=trial;                  #each trial has its own within-trial
                                        variance;
parms / parmsdata=BCG_data              #the parmsdata-option reads in the
                                        variable EST (indicating the within-
                                        trial variances) from the dataset
                                        BCG_data.sd2;
eqcons=1 to 13;                         #the within trial variances are
                                        considered to be known and must be kept
                                        constant;
run;
```

Running this analysis gives the following output:

```
                              The MIXED Procedure
(...)
                          Solution for Fixed Effects

Effect     Estimate    Std Error   DF       t  Pr >|t|  Alpha   Lower   Upper
INTERCEPT -0.43627138  0.04227521  12  -10.32  0.0001   0.05   -0.5284 -0.3442
```

The estimate of the common log odds ratio is equal to -0.436 with standard error = 0.042 leading to a 95% Wald based confidence interval of the log odds ratio from -0.519 to -0.353. (Although it seems overly precise, we will present results to three decimals, since these are used in further calculations and to facilitate comparisons between results of different models.) This corresponds to an estimate of 0.647 with a 95% confidence interval from 0.595 to 0.703 for the odds ratio itself. So we can conclude that vaccination is beneficial.

The confidence intervals and *p*-values provided by SAS Proc Mixed are based on the *t*-distribution rather than on the standard normal distribution, as is done in the standard likelihood approach. The number of degrees of freedom of the *t*-distribution is determined by Proc Mixed according to some algorithm. One can choose between several algorithms, but one can also specify in the model statement the number of degrees of freedom to be used for each covariable, except for the intercept. To get the standard Wald confidence interval and *p*-value for the intercept, the number of degrees of freedom used for the intercept should be specified to be ∞, which can be

accomplished by making a new intercept covariate equal to 1 and subsequently specifying 'no intercept' ('noint'). The SAS statement to be used is then:

```
model ln_or=int / s cl noint ddf=1000;
```

(the variable 'int' is a self-made intercept variable equal to 1).

Simple random effects model, maximum likelihood

The analysis under heterogeneity can be carried out by executing the following SAS statements. Unlike the previous model where we read in the within-trial variances from the datafile, we now specify the within trial variances explicitly in the 'parms'-statement. This has to be done because we want to define a grid of values for the first covariance parameter, i.e. the between trial variance, to get the profile likelihood function for the between trial variance to get its likelihood ratio based 95% confidence interval. Of course, one could also give only one starting value and read the data from a SAS-datafile like we did before.

 # THE RANDOM EFFECTS MODEL (MAXIMUM LIKELIHOOD);

```
Proc mixed cl method=ml data=BCG_data;

class trial;

model ln_or= / s cl;

random int/ subject=trial s;

repeated /group=trial;

parms (0.01 to 2.00 by 0.01)(0.35712)
(0.20813)(0.43341)(0.02031)(0.05195)
(0.00991)(0.22701)(0.00401)(0.05698)
(0.07542)(0.01253)(0.53416)(0.07164)
/eqcons=2 to 14;

make 'Parms' out=Parmsml;

run;
```

```
#call of procedure; 'cl' asks
for confidence intervals of
covariance parameters;
#trial is classification
variable;
#an intercept only model. print
fixed effect solution 's' and
its confidence limits 'cl';
#trial is specified as random
effect; 's' asks for the
empirical Bayes estimates;
#each trial has its own within
trial variance;
#defines grid of values for
between trial variance (from
0.01 to 1.00), followed by the
13 within trial variances which
are assumed to be known and must
be kept fixed;
#in the dataset 'Parms' the
maximum log likelihood for each
value of the grid specified for
the between trial variance is
stored, in order to read off the
profile likelihood based 95% CI
for the between trial variance;
```

Running this program gives the following output:

```
                              The MIXED Procedure

(...)
                    Covariance Parameter Estimates (MLE)

    Cov Parm    Subject  Group         Estimate  Alpha     Lower     Upper
    INTERCEPT   TRIAL                0.30245716   0.05    0.1350    1.1810

(...)

                           Solution for Fixed Effects

Effect     Estimate    Std Error   DF       t  Pr >|t|  Alpha   Lower   Upper
INTERCEPT -0.74197023  0.17795376  12   -4.17   0.0013  0.05   -1.1297 -0.3542
```

The ML-estimate of the mean log odds ratio is -0.742 with standard error 0.178. The standard Wald based 95% confidence interval is -1.091 to -0.393. (SAS Proc Mixed gives a slightly wider confidence interval based on a *t*-distribution with df=12). This corresponds to an estimated odds ratio of 0.476 with a 95% confidence interval from 0.336 to 0.675.

The ML-estimate of the between trial variance σ^2 is equal to 0.302. For each value of the grid specified in the 'Parms'-statement for the between trial variance (in the example the grid runs from 0.01 to 2.00 with steps of 0.01), the maximum log likelihood value is stored as variable 'LL' in the SAS-file '`Parmsml.sd2`'. Plotting the maximum log likelihood values against the grid of between trial variances gives the profile likelihood plot for the between trial variance presented in Figure 1. From this plot or a listing of the data set '`Parmsml.sd2`' one can read off the profile likelihood based 95% confidence interval for the between trial variance σ^2. This is done by looking for the two values of the between trial variance with a corresponding log likelihood of 1.92 lower than the maximum log likelihood. The 95% profile likelihood based confidence interval for σ^2 is (0.12, 0.89). (SAS Proc Mixed gives a Satterthwaite approximation based 95% confidence interval running from 0.135 to 1.180.)

Notice that by comparing the maximum log likelihood of this model with the previous fixed effects model, one gets the likelihood ratio test for homogeneity (the *p*-value has to be halved, because $\sigma^2 = 0$ is on the boundary of the parameter space).

A profile likelihood based confidence interval for the mean treatment effect ϑ can be made by trial and error by defining the variable `y=ln_or-c` as dependent variable for various values of `c` and specifying a model without intercept (add '`noint`' after the slash in the model statement). Then look for the two values of c that decrease the maximum log-likelihood by 1.92. The profile log-likelihood plot for ϑ is given in Figure 2.

The 95% confidence interval for the log odds ratio ϑ is (-1.13, -0.37), slightly wider

than the simple Wald approximation given above. This corresponds with a 95% confidence interval for the odds ratio of 0.323 to 0.691.

Remark: In Proc Mixed one can also choose the restricted maximum likelihood (REML) estimate (specify `method=reml` instead of `method=ml`). Then the resulting estimate for the between trial variance σ^2 is identical to the iterated DerSimonian-Laird estimator[5]. However, in this case the profile likelihood function should not be used to make a confidence interval for the log odds ratio ϑ. The reason is that differences between maximised REML likelihoods cannot be used to test hypotheses concerning fixed parameters in a general linear mixed model [20].

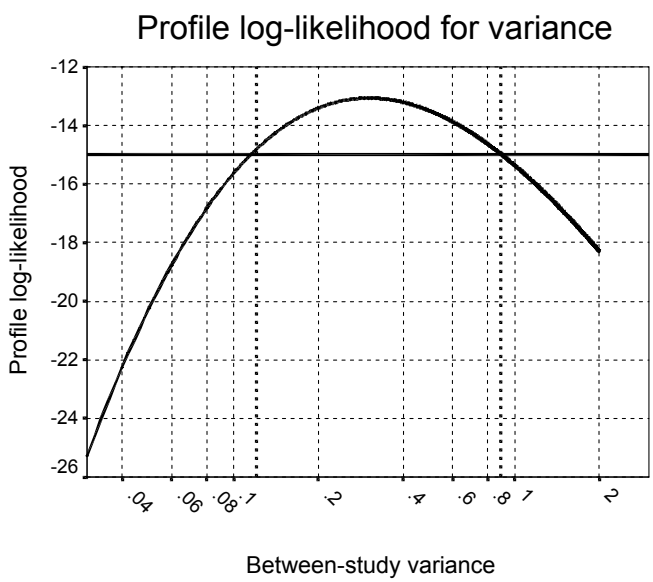


Figure 1. The 95% confidence interval of the between-trial variance σ^2 based on the profile likelihood function: (0.12 , 0.89).

The observed and corresponding empirical Bayes estimated log odds ratios with their 95% standard Wald respectively the 95% posterior confidence intervals per trial are presented in Figure 3. This figure shows the shrinkage of the empirical Bayes estimates towards the estimated mean log odds ratio and their corresponding smaller posterior confidence intervals .
The overall confidence interval of the mean true treatment effect and the overall prediction interval of the true treatment effect are given at the bottom of the figure.

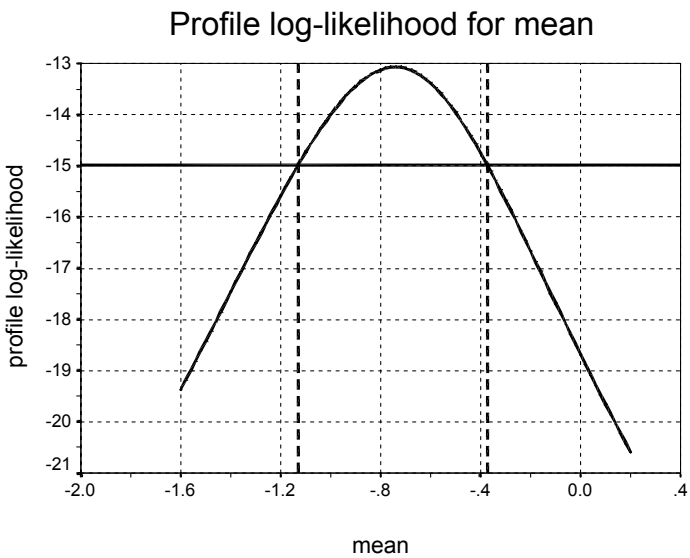


Figure 2. The 95% confidence interval of the treatment effect (log odds ratio) θ based on the profile likelihood function: (-1.13, -0.37).

The 95% prediction interval indicates the interval in which 95% of the true treatment effects of new trials are expected to fall. It is calculated as the ML-estimate plus and minus 1.96 times the estimated between trial standard deviation s and is here equal to (-1.820 to 0.336). The estimated probability for a new trial having a positive true treatment effect is $\Phi(0.742/0.302) = 0.993$.

4 Bivariate approach

In the previous section the parameter of interest was one-dimensional. In many situations it can be bivariate or even multivariate, for instance when there are more treatment groups or more outcome variables. In this section we discuss the case of a two-dimensional parameter of interest. We introduce the bivariate approach with special reference to the situation where one is interested in 'control rate regression', i.e. relating the treatment effect size to the risk of events in the control group. However, the approach applies generally.

Many studies show considerable variation in what is called the baseline risk. The baseline risk indicates the risk for patients under the control condition, which is the average risk of the patients in that trial when the patients were treated with the control treatment.

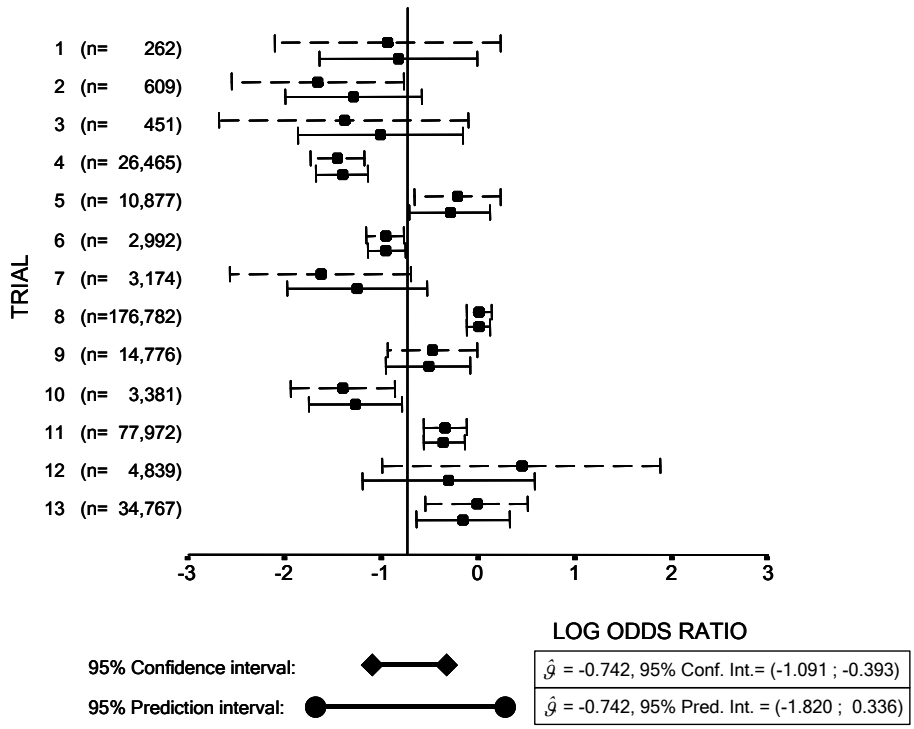


Figure 3. Forrest plot with the estimated log odds ratios of tuberculosis with their 95% confidence intervals in the trials included in the meta-analysis. The dashed horizontal lines indicate the standard Wald confidence intervals. The solid horizontal lines indicate the posterior or so-called empirical Bayes confidence intervals. The vertical line indicates the ML-estimate of the common (true) log odds ratio. Below the figure the 95% confidence interval for the mean log odds ratio and the 95% prediction interval for the true log odds ratio are presented.

One might wonder if there is a relation between treatment effect and baseline risk. Considering only the differences between the study arms may hide a lot information. Therefore, we think it is wise to consider the pair of outcomes of the two treatments. This is nicely done in the l'Abbé-plot[21], that gives a bivariate representation of the data by plotting the log odds in arm A versus the log odds in arm B. We show the plot in Figure 4 for the data of our example with A the vaccinated arm and B the not-vaccinated arm. The size of each circle represents the inverse of the variance of the log odds ratio in that trial. Points below the line of identity correspond to trials with an observed positive effect of vaccination.

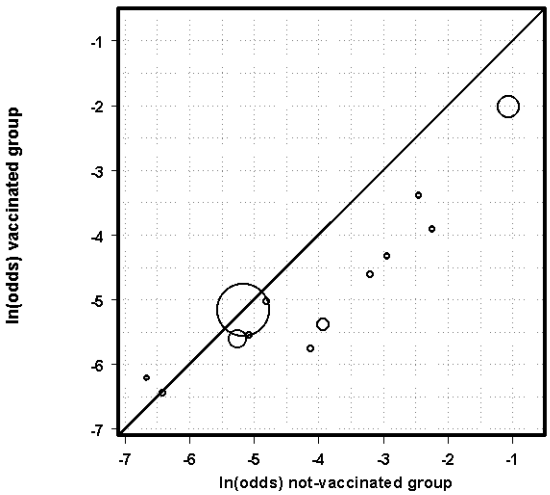


Figure 4. L'Abbé-plot of observed log(odds) of the not-vaccinated trial arm versus the vaccinated trial arm. The size of the circle is an indication for the inverse of the variance of the log odds ratio in that trial. Below the x=y line, the log odds in the vaccinated are lower than the log odds in the not-vaccinated arm, indicating that the vaccination works. On or above the x=y line, vaccination doesn't work beneficially.

The graph shows some effect of vaccination especially at the higher incidence rates.
A simple (approximate) bivariate model for any observed pair of arm specific outcome measures $\omega_i = (\hat{\omega}_{A,i}, \hat{\omega}_{B,i})$ with standard errors $(s_{A,i}, s_{B,i})$ in trial *i* is:

$$\begin{pmatrix} \hat{\omega}_{A,i} \\ \hat{\omega}_{B,i} \end{pmatrix} \sim N\left(\begin{pmatrix} \omega_{A,i} \\ \omega_{B,i} \end{pmatrix}, \begin{pmatrix} s^2_{A,i} & 0 \\ 0 & s^2_{B,i} \end{pmatrix} \right) \qquad (i = 1,...,n)$$

where $\omega_i = (\omega_{A,i}, \omega_{B,i})$ is the pair of true arm specific outcome measures for trial *i*. The conditional independence of $\hat{\omega}_A$ and $\hat{\omega}_B$ given the true ω_A and ω_B is a consequence of the randomised parallel study design and the fact that ω_A and ω_B are arm specific. In general, for instance in a cross-over study, or when ω_A and ω_B are treatment effects on two different outcome variables, the estimates might be correlated.
The mixed model approach assumes the pair $(\omega_{A,i}, \omega_{B,i})$ to follow a bivariate normal distribution, where, analogous to the univariate random effects model of section 3, the true outcome measures for both arms in the trials are normally distributed around some common mean treatment-arm outcome measure with a between-trial covariance matrix Σ:

$$\begin{pmatrix}\omega_{A,i}\\ \omega_{B,i}\end{pmatrix} \sim N(\begin{pmatrix}\omega_A\\ \omega_B\end{pmatrix}, \Sigma) \quad \text{with } \Sigma = \begin{pmatrix}\Sigma_{AA} & \Sigma_{AB}\\ \Sigma_{AB} & \Sigma_{BB}\end{pmatrix}$$

Σ_{AA} and Σ_{BB} describe the variability among trials in true risk under the vaccination and control condition, respectively. Σ_{AB} is the covariance between the true risk in vaccination and control group.

The resulting marginal model is

$$\begin{pmatrix}\hat{\omega}_{A,i}\\ \hat{\omega}_{B,i}\end{pmatrix} \sim N(\begin{pmatrix}\omega_A\\ \omega_B\end{pmatrix}, \Sigma + C_i)$$

with C_i the diagonal matrix with the s_i^2 's.

Maximum likelihood estimation for this model can be quite easily carried out by a self-made program based on the EM algorithm as the described in reference 2, but more practically convenient is to use appropriate mixed model software from statistical packages, such as the SAS procedure Proc Mixed.

Once the model is fitted, the following derived quantities are of interest:

- The mean difference $(\omega_A - \omega_B)$ and its standard error $\sqrt{(\text{var}(\omega_A) + \text{var}(\omega_B) - 2 \cdot \text{cov}(\omega_A, \omega_B))}$
- The population variance of the difference $\text{var}(\omega_A - \omega_B) = \Sigma_{AA} + \Sigma_{BB} - 2 \cdot \Sigma_{AB}$.
- The shape of the bivariate relation between the (true) ω_A and ω_B. That can be described by ellipses of equal density or by the regression lines of ω_A on ω_B and of the ω_B on ω_A. These lines can be obtained from classical bivariate normal theory. For example, the regression line of ω_A on ω_B has slope $\beta = \Sigma_{AB} / \Sigma_{BB}$ and residual variance $\Sigma_{AA} - \Sigma_{AB}^2 / \Sigma_{BB}$. The regression of the difference $(\omega_A - \omega_B)$ on either ω_A or ω_B can be derived similarly. At the end of this section we come back to the usefulness of these regression lines.

The standard errors of the regression slopes can be calculated from the covariance matrix of the estimated covariance parameters by the delta-method or by Fieller's method[22].

Example (continued): bivariate random effects model

As an example we carry out a bivariate meta-analysis with ω_A and ω_B the log odds of tuberculosis in the vaccinated and the not-vaccinated control arm, respectively. To execute a bivariate analysis in the SAS procedure Proc Mixed, we have to change the structure of the data set. Each treatment arm of a trial becomes a row in the data set, resulting in twice as many rows as in the original data set. The dependent variable is now the estimated log odds in a treatment arm instead of the log odds ratio. The new data set is called `BCGdata2.sd2` and the observed log odds is called `lno`. The standard error of the observed log odds, estimated by taking the square root of minus

the inverse of the second derivative of the log likelihood, is equal to $\sqrt{\frac{1}{x}+\frac{1}{n-x}}$, where n is the sample size of a treatment arm and x is the number of tuberculosis cases in a treatment arm. These standard errors are stored in the SAS data set `covvars2.sd2`.

The bivariate random effects analysis can be carried out by running the SAS commands given below. In the data step, the new dataset `BCGdata2.sd2` is made out of the dataset `BCGdata.sd2`, and the covariates are defined on trial arm level. The variable `exp` is 1 for the vaccinated (experimental) arms and 0 for the not-vaccinated (control) arms. The variable `con` is defined analogously with experimental and control conversed. The variable `arm` identifies the 26 unique treatment arms from the 13 studies, (here from 1 to 26); `latcon`, `latexp`, `yearcon` and `yearexp` are covariates to be used later. For numerical reasons we centralised the four variables `latcon`, `latexp`, `yearcon` and `yearexp` by subtracting the mean.

```
# The data step (bivariate analysis)
data bcgdata2;set bcg_data;
treat=1; lno=log(vd/vwd); var=1/vd+1/vwd; n=vd+vwd; output;
treat=0; lno=log(nvd/nvwd); var=1/nvd+1/nvwd; n=nvd+nvwd; output;
keep trial lno var n treat latitude--alloc;
run;
data bcgdata2;set bcgdata2;
arm=_n_; exp=(treat=1); con=(treat=0);
latcon=(treat=0)*(latitude-33); latexp=(treat=1)*(latitude-33);
yearcon=(treat=0)*(year-66); yearexp=(treat=1)*(year-66);
proc print noobs;run;
```

Running these SAS commands gives the following output:

TRIAL	LATITUDE	YEAR	ALLOC	TREAT	LNO	VAR	ARM	EXP	CON	LATCON	LATEXP	YEARCON	YEAREXP
1	44	48	1	1	-3.39283	0.25840	1	1	0	0	11	0	-18
1	44	48	1	0	-2.45413	0.09872	2	0	1	11	0	-18	0
2	55	49	1	1	-3.91202	0.17000	3	1	0	0	22	0	-17
2	55	49	1	0	-2.24583	0.03813	4	0	1	22	0	-17	0
3	42	60	1	1	-4.33073	0.33772	5	1	0	0	9	0	-6
3	42	60	1	0	-2.94444	0.09569	6	0	1	9	0	-6	0
4	52	77	1	1	-5.38597	0.01620	7	1	0	0	19	0	11

4	52	77	1	0	-3.92953	0.00411	8	0	1	19	0	11	0
5	13	73	2	1	-5.02786	0.03050	9	1	0	0	-20	0	7
5	13	73	2	0	-4.80872	0.02145	10	0	1	-20	0	7	0
6	44	53	2	1	-2.02302	0.00629	11	1	0	0	11	0	-13
6	44	53	2	0	-1.06490	0.00361	12	0	1	11	0	-13	0
7	19	73	1	1	-5.75930	0.12539	13	1	0	0	-14	0	7
7	19	73	1	0	-4.12552	0.10162	14	0	1	-14	0	7	0
8	13	80	1	1	-5.15924	0.00199	15	1	0	0	-20	0	14
8	13	80	1	0	-5.17126	0.00202	16	0	1	-20	0	14	0
9	27	68	1	1	-5.55135	0.03462	17	1	0	0	-6	0	2
9	27	68	1	0	-5.07961	0.02236	18	0	1	-6	0	2	0
10	42	61	3	1	-4.60458	0.05941	19	1	0	0	9	0	-5
10	42	61	3	0	-3.20337	0.01601	20	0	1	9	0	-5	0
11	18	74	3	1	-5.60295	0.00540	21	1	0	0	-15	0	8
11	18	74	3	0	-5.26210	0.00713	22	0	1	-15	0	8	0
12	33	69	3	1	-6.21180	0.20040	23	1	0	0	0	0	3
12	33	69	3	0	-6.65844	0.33376	24	0	1	0	0	3	0
13	33	76	3	1	-6.43840	0.03710	25	1	0	0	0	0	10
13	33	76	3	0	-6.42106	0.03454	26	0	1	0	0	10	0

```
# The procedure step (bivariate random effects analysis)
Proc mixed cl method=ml                  #call procedure; 'asycov' asks for
data=BCGdata2 asycov;                    asymptotic covariance matrix of
                                         covariance parameters
class trial arm;                         #trial and arm are classification
                                         variables;
model lno= exp con / noint s cl covb     #model with indicator variables 'exp'
ddf=1000, 1000;                          and 'con' as explanatory variables
                                         for log odds; confidence intervals
                                         and p-values for coefficients of
                                         'exp' and 'con' should be based on
                                         standard normal distribution (i.e. t-
                                         distribution with df = ∞).'covb'
                                         asks for covariance matrix of fixed
                                         effects parameters.
random exp con/ subject=trial            #experimental and control treatment
type=un s;                               are random effects, possibly
                                         correlated within a trial, and
                                         independent between trials;
                                         covariance matrix (Σ) is
                                         unstructured; print empirical Bayes
                                         estimates 's';
repeated /group=arm;                     #each study-arm in each trial has its
                                         own within study-arm variance (matrix
                                         Ci); within study estimation errors
                                         are independent (default);
estimate 'difference' exp 1 con -        #the 'estimate' command produces
1/cl df=1000;                            estimates of linear combinations of
                                         the fixed parameters with standard
```

```
                                        error computed from the covariance
                                        matrix of the estimates. Here we ask
                                        for the estimate of mean log odds
                                        ratio;
parms /parmsdata=covvars2 eqcons=4      #data file covvars2.sd2 contains the
to 29;                                  variable 'est' with starting values
                                        for the three covariance parameters
                                        of the random effects together with
                                        the 26 within study-arm variances.
                                        The latter are assumed to be known
                                        and should be kept fixed;
run;
```

Running this program gives the following output:

```
                          The MIXED Procedure
(...)
                  Covariance Parameter Estimates (MLE)

      Cov Parm  Subject  Group       Estimate  Alpha     Lower     Upper
      UN(1,1)   TRIAL              1.43137384   0.05    0.7369    3.8894
      UN(2,1)   TRIAL              1.75732532   0.05    0.3378    3.1768
      UN(2,2)   TRIAL              2.40732608   0.05    1.2486    6.4330
(...)
                       Solution for Fixed Effects

Effect     Estimate    Std Error   DF      t  Pr > |t|  Alpha    Lower    Upper
 EXP     -4.83374538   0.33961722 1000 -14.23   0.0001  0.05  -5.5002   -4.1673
 CON     -4.09597366   0.43469692 1000  -9.42   0.0001  0.05  -4.9490   -3.2430

                  Covariance Matrix for Fixed Effects

                Effect   Row          COL1          COL2
                EXP        1    0.11533985    0.13599767
                CON        2    0.13599767    0.18896142

(...)
                       ESTIMATE Statement Results

Parameter   Estimate    Std Error   DF   t   Pr > |t|   Alpha  Lower     Upper
difference -0.73777172  0.17973848 1000 -4.10 0.0001    0.05  -1.0905   -0.3851
```

The fixed parameter estimates $\hat{\omega} = (\hat{\omega}_A, \hat{\omega}_B)$ = (-4.834 , -4.096) represent the estimated mean log odds in the vaccinated and non-vaccinated group, respectively. The between trial estimated variance of the log odds is $\hat{\Sigma}_{AA}$ = 1.431 in the vaccinated groups and $\hat{\Sigma}_{BB}$ = 2.407 in the not-vaccinated groups. The between trial covariance is

estimated to be $\hat{\Sigma}_{AB} = 1.757$. So, the estimated correlation between the true vaccinated and true control log odds is $\hat{\Sigma}_{AB} / (\sqrt{\hat{\Sigma}_{AA}} \cdot \sqrt{\hat{\Sigma}_{BB}}) = 0.947$. The estimated covariance matrix for the ML-estimates $\hat{\omega}_B$ and $\hat{\omega}_A$ is

$\begin{pmatrix} \text{var}(\hat{\omega}_A) & \text{cov}(\hat{\omega}_A, \hat{\omega}_B) \\ \text{cov}(\hat{\omega}_B, \hat{\omega}_A) & \text{var}(\hat{\omega}_B) \end{pmatrix} = \begin{pmatrix} 0.115 & 0.136 \\ 0.136 & 0.189 \end{pmatrix}$. The estimated mean vaccination effect, measured as the log odds ratio, is equal to $(\hat{\omega}_A - \hat{\omega}_B) = (-4.834-(-4.096)) = -0.738$. The standard error of the mean vaccination effect is equal to $\sqrt{\text{var}(\hat{\omega}_A) + \text{var}(\hat{\omega}_B) - 2 \cdot \text{cov}(\hat{\omega}_A, \hat{\omega}_B)} = \sqrt{(0.115 + 0.189 - 2 \cdot 0.136)} = 0.180$, almost identical to the result of the univariate mixed model. This corresponds to an estimated odds ratio of exp(-0.738)=0.478 with a 95% confidence interval equal to (0.336; 0.680), again strongly suggesting an average beneficial vaccination effect. The slope of the regression line to predict the log odds in the vaccinated group from the log odds in the not-vaccinated group is equal to $\beta_{AB} = \hat{\Sigma}_{AB} / \hat{\Sigma}_{BB} = (1.757/2.407) = 0.730$. The slope of the reverse relationship is equal to $\beta_{BA} = \hat{\Sigma}_{AB} / \hat{\Sigma}_{AA} = (1.757/1.431) = 1.228$.

The variance of the treatment effect, measured as the log odds ratio, calculated from $\hat{\Sigma}$ is (1.431+2.407-2·1.757)=0.324, which is only slightly different from what we found earlier in the univariate random effects analysis. The conditional variance of the true log odds , and therefore also of the log odds ratio, in the vaccinated group given the true log odds in the not-vaccinated group is $(\Sigma_{AA} - \Sigma_{AB}^2 / \Sigma_{BB}) = (1.431-1.757^2/2.407) = 0.149$, which is interpreted as the variance between treatment effects among trials with the same baseline risk. So baseline risk, measured as the true log odds in the not-vaccinated group, explains (0.324-0.149)/ 0.324 = 54% of the heterogeneity in vaccination effect between the trials.

The 95% coverage region of the estimated bivariate distribution can be plotted in the so-called l'Abbé-plot[21] in Figure 5.

Figure 5 nicely shows that the vaccination effect depends on the baseline risk (log odds in not-vaccinated group) and that the heterogeneity in the difference between the log odds in the vaccinated versus the not-vaccinated treatment arms is for a large part explained by the regression coefficient being substantially smaller than 1. It also shows the shrinkage of the empirical Bayes estimates towards the main axis of the ellipse.

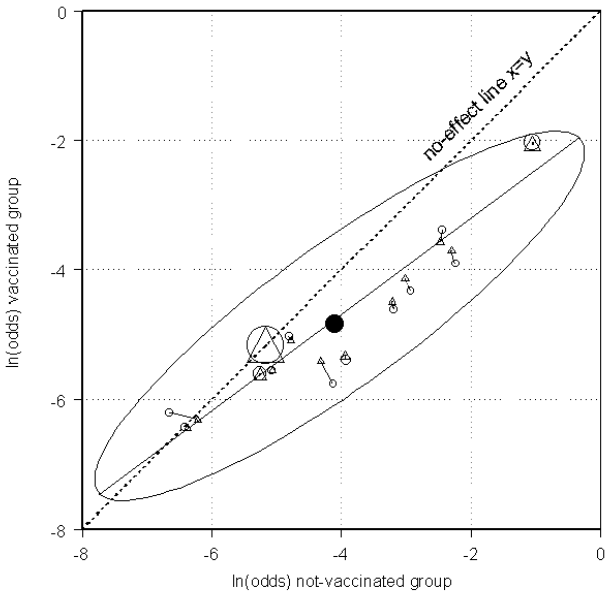


Figure 5. The 95% coverage region for the pairs of true log odds under vaccination and non-vaccination. The diagonal line is the line of equality between the two log odds.
Observed data from the trials are indicated with o, the empirical Bayes estimates are indicated with Δ. The common mean is indicated with the ● central in the plot. The ellipse is obtained from a line plot based on the equation $(x-\hat{\omega})\hat{\Sigma}^{-1}(x-\hat{\omega})' = 5.99$

In this example we specified the model in Proc Mixed as a model with two random intercepts, in which the fixed parameters correspond to ω_A and ω_B . An alternative would be to specify the model as a random-intercept-random-slope model, in which the fixed parameters correspond to ω_B and the mean treatment effect ω_A - ω_B. Then the SAS commands should be modified as follows:

```
model lno=treat/s cl covb ddf=1000;
random int treat/subject=trial type=un s;
```

Here `int` refers to a random trial specific intercept.

4.1 Relation between effect and baseline risk

The relation between treatment effect and baseline risk has been very much discussed in the literature[6-9, 23-30]. There are two issues that complicate the matter:

1. The relation between 'observed difference A-B' and 'observed baseline risk B' is prone to spurious correlation, since the measurement error in the latter is negatively correlated with measurement error in the first. It would be better to study B versus A or B-A versus (A+B)/2.

2. Even in the regression of 'observed risk in group A' on 'observed baseline risk in group B', which is not hampered by correlated measurement errors, the estimated slope is attenuated due to measurement error in the observed baseline risk[31].

See for an extensive discussion of these problems the article of Sharp et al.[32].

In dealing with measurement error there are two approaches[31, 33]

1. The 'functional equation' approach: true regressors as nuisance parameters.
2. The 'structural equation' approach: true regressors as random quantities with an unknown distribution.

The usual likelihood theory is not guaranteed to work for the functional equation approach because of the large number of nuisance parameters. The estimators may be inconsistent or have the wrong standard errors. The bivariate mixed model approach to meta-analysis used in this chapter is in the spirit of the structural approach. The likelihood method does work for the structural equation approach, so in this respect our approach is safe. Of course, the question of robustness of the results against misspecification of the mixing distribution is raised. However, Verbeke and Lesaffre[34] have shown that in the general linear mixed model the fixed effect parameters as well as the covariance parameters are still consistently estimated when the distribution of the random effects is misspecified, so long the covariance structure is correct. So our approach yields (asymptotically) unbiased estimates of slope and intercept of the regression line even if the normal distribution assumption is not fulfilled, although the standard errors might be wrong. Verbeke and Lesaffre[34] give a general method for robust estimation of the standard errors.

The mix of many fixed and a few random effects as proposed by Thompson et al.[8] and the models of Walter[9] and Cook and Walter[29] are more in the spirit of the functional approach. These methods are meant to impose no conditions on the distribution of the true baseline risks. The method of Walter[9] was criticised by Bernsen et al.[35]. Sharp and Thompson[30] use other arguments to show that Walter's method is seriously flawed. In a letter to the editor by Van Houwelingen & Senn[36] following the article of Thompson et al.[8] Van Houwelingen and Senn[36] argue that putting Bayesian priors on all nuisance parameters as done by Thompson et al. does not help solving the inconsistency problem. This view is also supported in the chapter on Bayesian methods in the book of Carroll et al.[31]. It would be interesting to apply the ideas of Carroll et al.[31] in the setting of meta-analysis, but that is beyond the scope of this chapter. Arends et al.[10] compare, in a number of examples, the approach of Thompson et al.[8] with the method presented here and the results were in line with the remarks of Van Houwelingen and Senn[36]. Sharp and Thompson[30], comparing the different approaches in a number of examples,

remark that whether or not assuming a distribution for the true baseline risks remains a debatable issue.
Arends et al.[10] also compared the approximate likelihood method as presented here with an exact likelihood approach where the parameters are estimated in a Bayesian manner with vague priors and found no relevant differences.

5 Meta-regression

In case of substantial heterogeneity between the studies, it is the statistician's duty to explore possible causes of the heterogeneity[15, 37-39]. In the context of meta-analysis that can be done by covariates on the study level that could 'explain' the differences between the studies. The term meta-regression to describe such analysis goes back to papers by Bashore et al.[40], Jones[41], Greenland[42] and Berlin and Antman[37]. We consider only analyses at the aggregated meta-analytic level. Aggregated information (mean age, percentage males) can describe the differences between studies. We will not go into covariates on the individual level. If such information exists, the data should be analysed on the individual patient level by hierarchical models. That is possible and a sensible thing to do, but beyond the scope of this chapter. We will also not consider covariates on the study arm level. That can be relevant in non-balanced observational studies. Such covariates could both correct the treatment-effect itself in case of confounding as well as explain existing heterogeneity between studies. Although the methods presented in this chapter might be applied straightforwardly, we will restrict attention to balanced studies in which no systematic difference between the study arms is expected.
Since the number of studies in a meta-analysis is usually quite small, there is a great danger of overfitting. The rule of thumb of one explanatory variable for each 5 (10) 'cases' leaves only room for a few explanatory variables in a meta-regression. In the example we have three covariates available: latitude, year of study and method of treatment allocation. Details are given in Table 1.
In the previous section we have seen that heterogeneity between studies can be partly explained by differences in baseline risk. So, it is also important to investigate whether covariates on the study level are associated with the baseline risk. That asks for a truly multivariate regression with a two-dimensional outcome, but we will start with the simpler regression for the one-dimensional treatment effect difference measure.

5.1 Regression for difference measure

Let X_i stand for the (row-)vector of covariates of study i including the constant term. Meta-regression relates the true difference ϑ_i to the 'predictor' $X_i\beta$. This relation cannot be expected to be perfect; there might be some residual heterogeneity that could be modelled by a normal distribution once again, that is $\vartheta_i \sim \mathrm{N}(X_i\beta, \sigma^2)$. Taking into account the imprecision of the observed difference measure $\hat{\vartheta}_i$ we get the marginal approximate model

$\hat{\vartheta}_i \sim \mathrm{N}(X_i\beta, \sigma^2 + s_i^2)$.

This model could be fitted by iteratively re-weighted least squares, where a new estimate of σ^2 is used in each iteration step or by full maximum likelihood with appropriate software. In the sequel we will describe how the model can be fitted by SAS.

Example (continued)

A graphical presentation of the data is given in Figure 6. Latitude and year of publication both seem to be associated with the log odds ratio, while latitude and year are also correlated. Furthermore, at first sight, the three forms of allocation seem to have little different average treatment effects.

Regression on latitude

The regression analysis for the log odds ratio on latitude can be carried out by running the following mixed model in SAS:

```
Proc mixed cl method=ml                 #call procedure;
data=BCG_data;
class trial;                            #trial is classification variable;
model ln_or= latitude / s cl            #latitude is only predictor variable;
covb;
random int/ subject=trial s;            #random trial effect;
repeated /group=trial;                  #each trial has its own within study
                                        variances;
parms /parmsdata=covvars3               #data set covvars3 contains a starting
eqcons=2 to 14;                         value for between study variance and
                                        13 within study variances which should
                                        be kept fixed;
run;
```

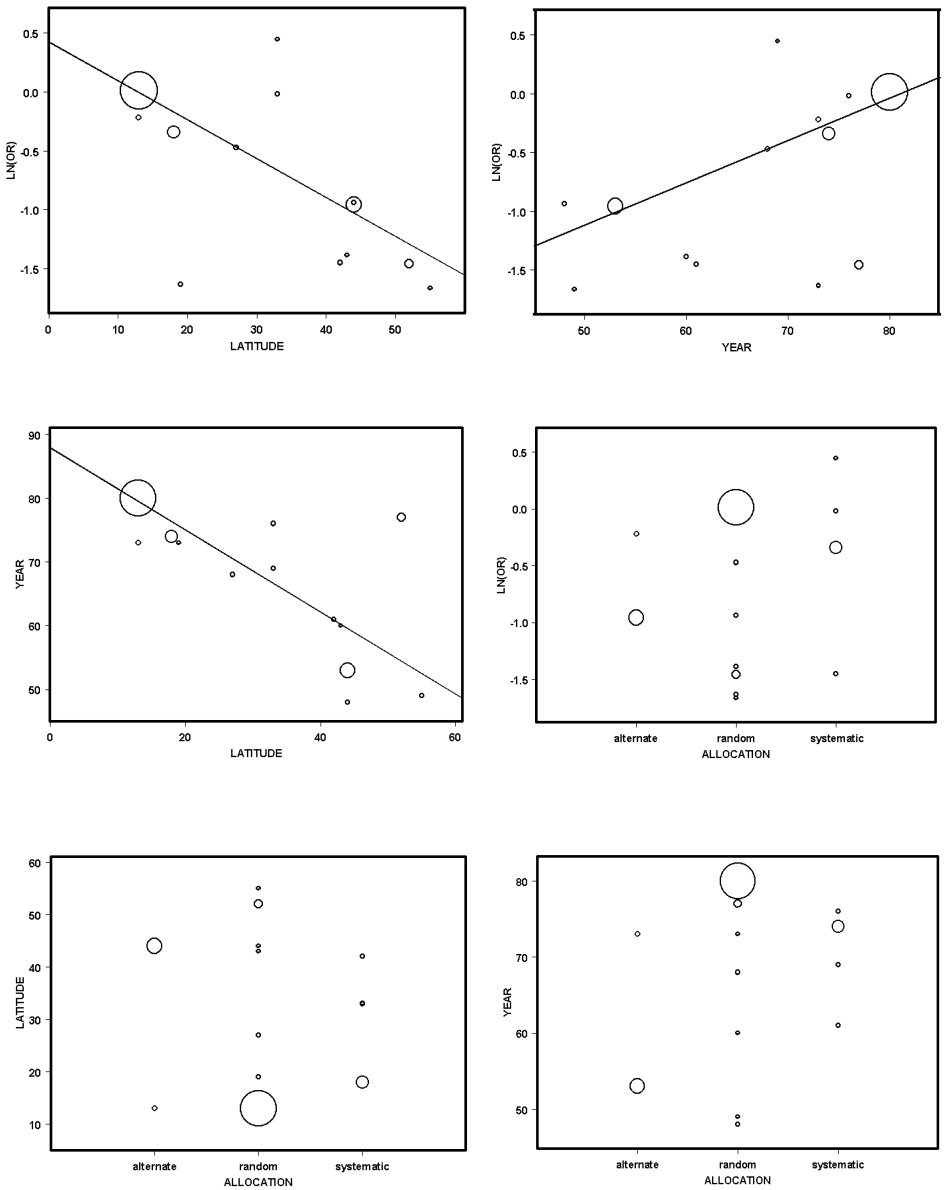


Figure 6. Graphical relationships between the variables with a weighted least squares regression line. The size of the circle corresponds to the inverse variance of the log odds ratio in that trial.

Running this program gives the following output:

```
                          The MIXED Procedure

(...)              Covariance Parameter Estimates (MLE)
    Cov Parm    Subject  Group         Estimate  Alpha     Lower     Upper
    INTERCEPT   TRIAL               0.00399452   0.05    0.0004  1.616E29

(...)                   Solution for Fixed Effects

Effect       Estimate    Std Error   DF    t     Pr > |t| Alpha  Lower  Upper
INTERCEPT    0.37108745  0.10596655  11    3.50  0.0050   0.05   0.1379 0.6043
LATITUDE    -0.03272329  0.00337134   0   -9.71        .   0.05        .      .

                    Covariance Matrix for Fixed Effects

                  Effect      Row          COL1          COL2
                  INTERCEPT     1    0.01122891   -0.00031190
                  LATITUDE      2   -0.00031190    0.00001137
```

The residual between study variance in this analysis turns out to be 0.004, which is dramatically smaller than the between study variance of 0.302 in the random effect model above without the covariate latitude in the model. So latitude explains 98.7% of the between trials variance in treatment effects differences. The regression coefficients for the intercept and for latitude are 0.371 (standard error = 0.106) and -0.033 (standard error=0.003), respectively. The estimated correlation between these estimated regression coefficients is –0.873.

Just for comparison we give the results of an ordinary weighted linear regression. The weights are equal to the inverse squared standard error of the log odds ratio, instead of the correct weights equal to the inverse squared standard error of the log odds ratio plus $\hat{\sigma}^2$. The intercept was 0.395 (se = 0.124) and the slope -0.033 (se = 0.004). The results are only slightly different, which is explained by the very small residual between study variance.

Regression on year

Running the same model as above with only changing latitude into year, the residual between study variance becomes 0.209. So year of publication explains 30.8% of the between trials variance in treatment effects differences, much less than the variance explained by the covariate latitude. The regression coefficients for the intercept and for year are -2.800 (standard error = 1.031) and 0.030 (standard error = 0.015), respectively. The estimated correlation between these estimated regression coefficients is -0.989.

Again, just for comparison, we also give the results of the ordinary weighted linear regression. The intercept was –2.842 (se = 0.876) and the slope 0.033 (se = 0.012). Like in the previous example, the differences are relatively small.

Regression on allocation

Running the model with allocation as only (categorical) covariate (in the SAS commands, specify: `class trial alloc;`), gives a residual between study variance equal to 0.281. This means that only 7% of the between trial variance in the treatment effect differences is explained by the different forms of allocation. The treatment effects (log odds ratio) do not differ significantly between the trials with random, alternate and systematic allocation
(p= 0.396).

Regression on latitude and year

When both covariates latitude and year are put into the model the residual between studies variance becomes only 0.002, corresponding with an explained variance of 99.3%, only slightly more than by latitude alone. The regression coefficients for the intercept, latitude and year are respectively 0.494 (standard error = 0.529), -0.034 (standard error = 0.004) and -0.001 (standard error = 0.006).

We conclude that latitude gives the best explanation of the differences in vaccination effect between the trials, since it already explains 98% of the variation. Since the residual variance is so small, the regression equation in this example could have been obtained by ordinary weighted linear regression under the assumption of homogeneity.

In the original medical report[4] on this meta-analysis the authors mentioned the strong relationship between treatment effect and latitude as well. They speculated that the biological explanation might be the presence of nontuberculous myobacteria in the population, which is associated with geographical latitude.

Goodness-of-fit of the model obtained above can be checked as in the weighted least squares approach by individual standardisation of the residuals $(\hat{\vartheta}_i - X_i\hat{\beta})/\sqrt{\sigma^2 + s_i^2}$ and using standard goodness-of-fit checks.

In interpreting the results of meta-regression analysis, it should be kept in mind that this is all completely observational. Clinical judgement is essential for correct understanding of what is going on. Baseline risk may be an important confounder and we will study its effect below.

5.2 Bivariate regression

The basis of the model is the relation between the pair $(\omega_{A,i}, \omega_{B,i})$, for example (true log odds in vaccinated group, true log odds in control group), and the covariate vector X_i. Since the covariate has influence on both components we have a truly multivariate regression problem in the classical sense, that can be modelled as

$$\begin{pmatrix} \omega_{A,i} \\ \omega_{B,i} \end{pmatrix} \sim N(BX_i, \Sigma)$$

Here, the matrix B is a matrix of regression coefficients: the first row for the A-component and the second row for the B-component. Taking into account the errors in the estimates we get the (approximate) model

$$\begin{pmatrix} \hat{\omega}_{A,i} \\ \hat{\omega}_{B,i} \end{pmatrix} \sim N(BX_i, \Sigma + C_i)$$

Fitting this model to the data can again be done by a self made program using the EM-algorithm or by programs as SAS Proc Mixed. The hardest part is the interpretation of the model. We will discuss the interpretation for the example.
So far we have shown for our leading example the univariate fixed effects model, the univariate random effect without covariates, the bivariate random effects model without covariates and eventually the univariate random effects model with covariates. We end this paragraph with a bivariate random effects model with covariates.

Example (continued): bivariate meta-analysis with covariates

To carry out the bivariate regression analyses in SAS Proc Mixed we need again the data set `BCGdata2.sd2` which was organised on treatment arm level. In this example we take latitude as covariate. The model can be fitted using the SAS code given below, where the variables `exp`, `con` and `arm` have the same meaning as in the bivariate analysis above without covariates. The variable `latcon` is for the not-vaccinated (control) groups equal to the latitude value of the trial and zero for the vaccinated (experimental) groups. The variable `latexp`, is defined analogously with vaccinated and non-vaccinated reversed.

```
Proc mixed cl method=ml data=BCGdata2;   #call procedure;
class trial arm;                         #trial and treatment arm are
                                         defined as classification
                                         variables;
model lno= con exp latcon latexp/noint   #model with indicator
s cl ddf=1000,1000,1000,1000;            variables 'exp' and 'con'
                                         together with latitude as
                                         explanatory variable for log
```

```
                                        odds in both treatment groups;
random con exp / subject=trial          #control arm and experimental
type=fa0(2) s ;                         trial arm are specified as
                                        random effects; covariance
                                        matrix is unstructured,
                                        parameterized as factor
                                        analytic;
repeated /group=arm;                    #each study-arm in each trial
                                        has its own within study-arm
                                        error variance;
parms /parmsdata=covvars4 eqcons=4 to   #in the data file covvars4
29;                                     three starting values are
                                        given for the between study
                                        covariance matrix, together
                                        with the 26 within study-arm
                                        variances. The latter are
                                        assumed to be known and kept
                                        fixed;
estimate 'difference slopes' latexp 1   # estimate of the difference
latcon -1                               in slope between the
/cl df=1000;                            vaccinated and not-vaccinated
                                        groups;
run;
```

Remark: In the program above we specified `type=fa0(2)`instead of `type=un` for Σ. If one chooses the latter, the covariance matrix is parameterized as

$$\begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3 \end{bmatrix}$$

and unfortunately the program does not converge if the estimated correlation is (very near to) 1, as is the case here. If one chooses the former, the covariance matrix is parameterized as

$$\begin{bmatrix} \alpha_{11}^2 & \alpha_{11}\alpha_{12} \\ \alpha_{11}\alpha_{12} & \alpha_{12}^2 + \alpha_{22}^2 \end{bmatrix}$$

and the program converges even if the estimated correlation is 1, i.e. if α_{22}=0.

Running the program gives the following output:

```
                          The MIXED Procedure
(...)              Covariance Parameter Estimates (MLE)
      Cov Parm  Subject  Group       Estimate  Alpha     Lower     Upper
      FA(1,1)   TRIAL              1.08715174   0.05    0.7582    1.6896
      FA(2,1)   TRIAL              1.10733154   0.05    0.6681    1.5466
      FA(2,2)   TRIAL             -0.00000000      .         .         .

(...)
```

Solution for Fixed Effects

Effect	Estimate	Std Error	DF	t	Pr > |t|	Alpha	Lower	Upper
CON	-4.11736845	0.30605608	1000	-13.45	0.0001	0.05	-4.7180	-3.5168
EXP	-4.82570990	0.31287126	1000	-15.42	0.0001	0.05	-5.4397	-4.2118
LATCON	0.07246261	0.02192060	1000	3.31	0.0010	0.05	0.0294	0.1155
LATEXP	0.03913388	0.02239960	1000	1.75	0.0809	0.05	-0.0048	0.0831

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DF	t	Pr > |t|	Alpha	Lower	Upper
difference slopes	-0.03332874	0.00284902	1000	-11.70	0.0001	0.05	-0.0389	-0.0277

In Figure 7 the relationship between latitude and the log odds of tuberculosis is presented for the vaccinated treatment arms A as well as for the non-vaccinated treatment arms B. For the not-vaccinated trial arms the regression line is log(odds)= -0.4117 + 0.072· (latitude - 33) = -6.509 + 0.072· latitude (standard errors of intercept and slope are 0.794 and 0.022, respectively). Notice that latitude was centralised at latitude=33 (see page 64). For the vaccinated trial arms the regression line is log(odds) = -0.483· (latitude - 33) = -6.117 + 0.039· latitude (standard errors of intercept and slope are 0.809 and 0.022, respectively). We see that latitude has a strong effect, especially on the log odds of the non-vaccinated study group.

The between study covariance matrix $\hat{\Sigma}$ is equal to the nearly singular matrix

$$\begin{bmatrix} 1.1819 & 1.2038 \\ 1.2038 & 1.2262 \end{bmatrix}$$

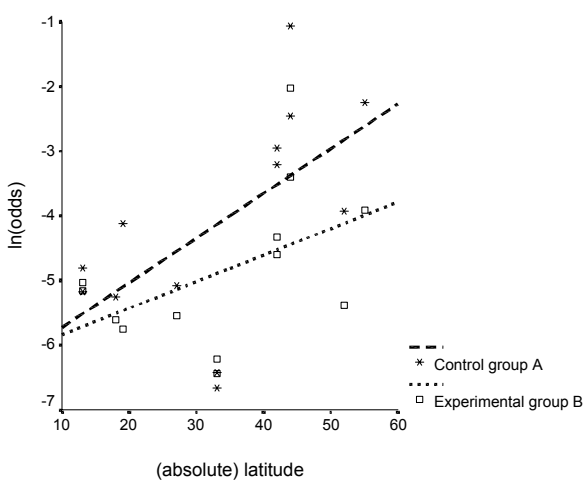


Figure 7. Log odds versus latitude for control group A and experimental group B

The estimated regression line of the treatment difference measure on latitude is:
log odds ratio$_{A\ vs\ B}$ = 0.392 - 0.033·latitude, with standard errors 0.093 and 0.003 for intercept and slope, respectively. This regression line is almost identical to the one resulting from the univariate analysis in the previous example. The estimated residual between study variance is only 0.0003, meaning that latitude explains almost all heterogeneity in the treatment effects.
The regression line of the difference measure on both latitude and baseline risk is:
log odds ratio$_{A\ vs\ B}$ = 0.512 - 0.039· latitude+ 0.019· log odds$_B$.
The standard errors can be calculated by the delta method. We see that the regression coefficient of the baseline log odds is quite small compared to the analysis without any covariates.
The results of this bivariate regression and the results of the simple bivariate model without covariates of section 4 are summarised in Table 2. By explaining variation in treatment effects by latitude, hardly any residual variation is left. Although this is all observational, we come to the tentative conclusion that the effect of vaccination depends on latitude rather than on baseline risk.

Table 2. Residual variance of treatment effect in different meta-regression models.

Explanatory variables in the model	Residual variance of treatment effect
No covariates	0.324
Baseline	0.149
Latitude	0.0003
Baseline + Latitude	0.0001

6 Extensions: exact likelihoods, non-normal mixtures, multiple endpoints

The approximate likelihood solutions may be suspected if the sample sizes per study are relative small. There are different approaches to repair this and to make the likelihoods less approximate. We will first discuss the bivariate analysis where things are relatively easy and then the analysis of difference measures.

6.1 More precise analysis of bivariate data

Here, the outcome measures per study arm are direct Maximum Likelihood Estimates of the relevant parameter. The estimated standard error is derived from the second derivative of the log-likelihood evaluated at the ML-estimate. Our approach is an approximation for fitting a generalised linear mixed model (GLMM) by the maximum likelihood method. The latter is hard to carry out. A popular approximation is by means of the second order Laplace approximation or the equivalent PQL-method[43], that is based on an iterative scheme where the second derivative is evaluated at the posterior mode. This can easily be mimicked in the SAS procedure Proc Mixed by iteratively replacing the estimated standard error computed from the empirical Bayes estimate as yielded by the software. For the analysis of log odds's as in the example, one should realise that the variance of log odds is derived from the second derivative of the log-likelihood evaluated at the ML-estimate of p, and is given by $1/(np(1-p))$. In the first iteration, p is estimated by the fraction of events in the study arm. In the next iteration p is replaced by the value derived from the empirical Bayes estimate for log odds. This is not very hard to do and easy to implement in a SAS macro that iteratively uses Proc Mixed (see the example below, the macro version is available from the authors).
This will help for intermediate sample sizes and moderate random effect variances. There are however situations thinkable (small samples, large random effect variances) in which the second order approximations do not work[44] and one has to be very careful in computing and maximising the likelihood's. Fortunately, that is much more of a problem for random effects at the individual level than at the aggregated level we have here.

Example (continued)

After running the bivariate random effects model discussed in section 4, the empirical Bayes estimates can be saved by adding the statement:

```
make 'Predicted' out=Pred;
```

in the Proc Mixed command and adding a '`p`' after the slash in the model statement. In this way the empirical Bayes estimates for log odds are stored as variable `_PRED_` in the new SAS data-file `Pred.sd2`. The within-trial variances in the next iteration of the SAS procedure Proc Mixed are derived from these empirical Bayes estimates in the way we described above. The three starting values needed for the between trial variance matrix are stored as variable `est` in the SAS -file `covvars5.sd2`.
So, after running the bivariate random effects model once and saving the empirical Bayes estimates for log odds, one can run the two data steps described below to compute the new estimates for the within-trial variances, use these within-trial

variances in the next bivariate mixed model, save the new empirical Bayes estimates and repeat the whole loop. This iterative process should be continued until the parameter estimates converge.

```
# Data step to combine empirical Bayes estimates and original datafile
from section 4 and to calculate the new within-trial variances;
data Pred1;
merge BCGdata2 Pred;
pi=exp(_PRED_)/(1+exp(_PRED_));
est=1/(n*pi*(1-pi));
run;

# Data step to create the total datafile that is needed in the Parms-
statement (between- and within-trial variances);
data Pred2;
set covvars5 Pred1;
run;

# Procedure step to run the bivariate random effects model with new
within-trial variances, based on the empirical Bayes estimates.
proc mixed cl method=ml data=BCGdata2 asycov;
class trial arm;
model lno= exp con / p noint s cl covb ddf=1000, 1000;
random exp con/ subject=trial type=un s;
repeated /group=arm subject=arm;
estimate 'difference' exp 1 con -1 / cl df=1000;
parms / parmsdata=Pred2 eqcons=4 to 29;
run;
```

Running the data steps and the mixed model iteratively until convergence is reached, gives the following output:

```
                          The MIXED Procedure
(...)
                Covariance Parameter Estimates (MLE)
      Cov Parm  Subject  Group     Estimate  Alpha     Lower     Upper
      UN(1,1)   TRIAL            1.43655989   0.05    0.7392    3.9084
      UN(2,1)   TRIAL            1.76956270   0.05    0.3395    3.1996
      UN(2,2)   TRIAL            2.43849037   0.05    1.2663    6.4991

                       Solution for Fixed Effects

Effect   Estimate    Std Error   DF      t    Pr > |t|  Alpha   Lower    Upper

EXP     -4.84981269  0.34001654  1000  -14.26  0.0001   0.05   -5.5170  -4.1826
CON     -4.10942999  0.43736103  1000   -9.40  0.0001   0.05   -4.9677  -3.2512
```

Covariance Matrix for Fixed Effects

Effect	Row	COL1	COL2
EXP	1	0.11561125	0.13690215
CON	2	0.13690215	0.19128467

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DF	t	Pr > |t|	Alpha	Lower	Upper
difference	-0.74038270	0.18191102	1000	-4.07	0.0001	0.05	-1.0974	-0.3834

The mean outcome measures (log odds) for arms A and B are, respectively, -4.850 (standard error = 0.340) and -4.109 (standard error = 0.437). The between trial variance of the log odds in the vaccinated treatment arm A is $\hat{\Sigma}_{AA} = 1.437$ and $\hat{\Sigma}_{BB} = 2.438$ in the not-vaccinated arm B. The estimate of the between trial covariance is equal to $\hat{\Sigma}_{AB} = 1.770$. The estimated mean vaccination effect in terms of the log odds ratio is -0.740 (standard error = 0.182). In this example, convergence was already reached after one or two iterations. The final estimates are very similar to the original bivariate random effects analysis we have discussed in section 4, where the mean outcome measures $\hat{\omega}_A$ and $\hat{\omega}_B$ were respectively -4.834 (s.e.=0.340) and -4.096 (s.e.= 0.434) . Of course, when the number of patients in the trials were smaller, the benefit and necessity of this method would be more substantial.

Another possibility if the approximate likelihood solutions are suspected is to use the exact likelihood, based on the binomial distribution of the number of events per treatment arm, and to estimate the parameters following a Bayesian approach with vague priors in combination with Markov Chain Monte Carlo (MCMC) methods[45]. Arends et al.[10] give examples of this approach. In their examples the difference with the approximate likelihood estimates turned out to be very small.

6.2 More precise analysis of difference measures

The analysis of difference measures, i.e. one summary measure per trial characterising the difference in efficacy between treatments, is a bit more complicated because the baseline value is considered to be a nuisance parameter. Having this nuisance parameter can be avoided and a lot of 'exactness' in the analysis can be gained by suitable conditioning on ancillary statistics. In the case of binary outcomes one can condition on the marginals of the 2×2-tables and end up with the non-central hypergeometric distribution that only depends on the log odds ratio. Details are given in Van Houwelingen et al.[2].

However, the hypergeometric distribution is far from easy to handle and it does not seem very attractive to try to incorporate covariates in such an analysis as well. The

bivariate analysis is much easier to carry out at the price of the assumption that the baseline parameter follows a normal distribution. However, that assumption can be relaxed as well and brings us to the next extension: the non-normal mixture.

6.3 Non-Normal Mixtures

The assumption of a normal distribution for the random effects might not be realistic. Technically speaking it is not very hard to replace the normal mixture by a fully non-parametric mixture. As is shown by Laird[46] the Nonparametric Maximum Likelihood Estimator of the mixing distribution is always a discrete mixture and can easily be estimated by means of the EM algorithm[47]. An alternative is to use the software C.A.MAN of Böhning et al.[48]. However, just fitting a completely nonparametric mixture is no good way of checking the plausibility of the normal mixture. The nonparametric estimates are always very discrete even if the true mixture is normal. A better way is to see whether a mixture of two normals (with the same variance) fits better than a single normal. This model can describe a very broad class of distributions: unimodal as well as bimodal, symmetric as well as very skewed [19]. Another way is to estimate the skewness of the mixture somehow and mistrust the normality if the skewness is too big. It should be realised however, that estimating mixtures is a kind of ill-posed problem and reliable estimates are hard to obtain[49]. To give an impression we fitted a nonparametric mixture with the homemade program based on the EM algorithm described in Van Houwelingen et al.[2] to the log odds ratio of our example using approximate likelihoods.

Result:	atom	probability
	-1.4577	0.3552
	-0.9678	0.1505
	-0.3296	0.2980
	0.0023	0.1963

corresponding mean: -0.761

corresponding variance: 0.349

The first two moments agree quite well with the normal mixture. It is very hard to tell whether this 4-point mixture gives any evidence against normality of the mixture.

The bivariate normal mixture of section 4 is even harder to check. Non-parametric mixtures are hard to fit in two dimensions. An interesting question is whether the estimated regression slopes are robust against non-normality. Arends et al.[10] modelled the baseline distribution with a mixture of two normal distributions and found in all their examples a negligible difference with modelling the baseline parameter with one normal distribution, indicating that the method is robust indeed[10]. However, this was only based on three examples and we do not exclude

the possibility that in some other data examples the regression slopes might be more different.

6.4 Multiple outcomes

In a recent paper Berkey et al.[50] discuss a meta-analysis with multiple outcomes. A similar model is used in the context of meta-analysis of surrogate markers by Daniels and Hughes[51] and discussed by Gail et al.[52]. In the simplest case of treatment difference measures for several outcomes, the situation is very similar to the bivariate analysis of sections 4 and 5. The model $\begin{pmatrix} \omega_{\mathrm{A},i} \\ \omega_{\mathrm{B},i} \end{pmatrix} \sim \mathrm{N}(BX_i, \Sigma)$ could be used, where ω_A stands for the (difference) measure on outcome A and ω_B for the measure on outcome B. It could easily be generalised to more measures C, D, etc. The main difference is that the estimated effects are now obtained in the same sample and, therefore, will be correlated. An estimate of this correlation is needed to perform the analysis. The only thing that changes in comparison with section 5 is that the matrix C_i in $\begin{pmatrix} \hat{\omega}_{\mathrm{A},i} \\ \hat{\omega}_{\mathrm{B},i} \end{pmatrix} \sim \mathrm{N}(BX_i, \Sigma + C_i)$ is not diagonal anymore but allows within-trial covariation.

This approach can easily be adapted to the situation where there more than two outcome variables or more treatment groups.

Example Berkey et al.[50]

Berkey et al.[50] illustrate several fixed and random (multivariate) meta-regression models using a meta-analysis from Antczak-Bouckoms et al.[53]. This meta-analysis concerns five randomised controlled trials, where a surgical procedure is compared with a non-surgical procedure. Per patient two outcomes are assessed: (pre- and post-treatment change in) probing depth (PD) and (pre- and post-treatment change in) attachment level (AL). Since the efficacy of the surgical procedure may improve over time, a potential factor that may influence the trial results is the year of publication[50]. The two treatment effect measures are defined as:

ω_{PD} = mean PD under surgical treatment - mean PD under non-surgical treatment

ω_{AL} = mean AL under surgical treatment - mean AL under non-surgical treatment

The data are given in Table 3.

Table 3. Data from the meta-analysis of Antczak-Bouckoms et al.[53]

trial	publication year	$\hat{\omega}_{\mathrm{PD},i}$	$\hat{\omega}_{\mathrm{AL},i}$	$\mathrm{var}(\hat{\omega}_{\mathrm{PD},i})$	$\mathrm{var}(\hat{\omega}_{\mathrm{AL},i})$	$\mathrm{cov}\,\mathrm{ar}(\hat{\omega}_{\mathrm{PD},i},\hat{\omega}_{\mathrm{AL},i})$
1	1983	0.47	-0.32	0.0075	0.0077	0.0030
2	1982	0.20	-0.60	0.0057	0.0008	0.0009
3	1979	0.40	-0.12	0.0021	0.0014	0.0007
4	1987	0.26	-0.31	0.0029	0.0015	0.0009
5	1988	0.56	-0.39	0.0148	0.0304	0.0072

As an example we fit the model with year of publication as explanatory variable. Berkey et al.[50] fitted this model using a self written program in SAS Proc IML. We show how it can be done with SAS Proc Mixed. The data setup is the same as in the earlier discussed bivariate models with two data rows per trial, one for each outcome measure. Also the Proc Mixed program is completely analogous. The only difference is that in the data set containing the elements of the C_i's now the covariance between the two outcomes per trial must be specified as well. The SAS code is:

```
proc mixed cl method=ml                  # call procedure;
data=berkey;
class trial type;                        # trial and outcome type (PD or AL)
                                         are classification variables;
model outcome=  pd al pdyear             # model with indicator variables 'pd'
alyear/noint s cl;                       and 'al' together with publication
                                         year as explanatory variable;
random pd al / subject=trial             # specification of among-trial
type=un s ;                              covariance matrix for both outcomes;
repeated type /subject=trial             # specification of (non-diagonal)
group=trial type=un;                     within-trial covariance matrix;
parms /parmsdata=covvars6 eqcons=4       # covvars6 contains: 3 starting
to 18;                                   values for the two between trial
                                         variances and covariance, 10 within
                                         trial variances (5 per outcome
                                         measure) and 5 covariances. The last
                                         15 parameters are assumed to be known
                                         and must be kept fixed.
run;
```

Part of the SAS Proc Mixed output is given on the next page.

```
                    The MIXED Procedure
(...)
            Covariance Parameter Estimates (MLE)

      Cov Parm  Subject  Group      Estimate    Alpha    Lower     Upper

      UN(1,1)   TRIAL               0.00804054   0.05    0.0018    2.0771
      UN(2,1)   TRIAL               0.00934132   0.05   -0.0113    0.0300
      UN(2,2)   TRIAL               0.02501344   0.05    0.0092    0.1857

(...)
              Solution for Fixed Effects

Effect   Estimate     Std Error     DF   t    Pr > |t| Alpha  Lower   Upper
PD        0.34867848  0.05229098     3  6.67  0.0069  0.05   0.1823  0.5151
AL       -0.34379097  0.07912671     3 -4.34  0.0225  0.05  -0.5956 -0.0920
PDYEAR    0.00097466  0.01543690     0  0.06       .  0.05        .       .
ALYEAR   -0.01082781  0.02432860     0 -0.45       .  0.05        .       .
```

The estimated model is:

ω_{PD} = 0.34887 + 0.00097*(year-1984)

ω_{AL} = -0.34595 - 0.01082*(year-1984)

The standard errors of the slopes are 0.0154 and 0.0243 for PD and AL, respectively.

The estimated among-trial covariance matrix is:

$$\hat{\Sigma} = \begin{pmatrix} 0.008 & 0.009 \\ 0.009 & 0.025 \end{pmatrix}$$

The results are identical to those of Berkey et al.[50] with the random-effects multiple outcomes that were estimated with the method called by Berkey the Multivariate Maximum Likelihood (MML) method.

6.5 Other outcome measures

Our presentation concentrates on dichotomous outcomes. Much of it carries over to other effect measures that are measured on a different scale. For instance, our methods apply if the outcome variable is continuous and an estimate of the average outcome and its standard error is available in both treatment arms. However, in some cases only a relative effect is available, such as the standardized effect measure (difference in outcome/ standard deviation of the measurements in the control group) which is popular in psychological studies. In that case only the one-dimensional analysis applies. A special case is survival analysis. The log hazard ratio in the Cox model cannot be written as the difference of two effect measures. However, some measure of baseline risk, e.g. one-year survival rate in the control arm, might be defined and the bivariate outcome analysis described above can be used to explore

the relation between treatment effect and baseline risk. A complicating factor is that the two measures are not independent any more. However, if an estimate of the correlation between the two measures is available, the method can be applied.

6.6 Other software

Although we illustrated all our examples with the SAS procedure Proc Mixed, most if not all analyses could be carried out by other (general) statistical packages as well. A nice review of available software for meta-analysis is recently written by Sutton[54].
Any package like SPSS, SAS, S-Plus and Stata that can perform a weighted linear regression and suffices to perform a standard fixed-effect meta-analysis or a fixed effects meta-regression.
For fitting random effects models with approximate likelihood, a program for the General Linear Mixed Model (GLMM) is needed, which is available in many statistical packages. However not all GLMM programs are appropriate. One essential requirement of the program is that one can fix the within trial variance in the model at arbitrary values per trial.
In S-Plus the function *lme* is used to fit linear mixed effects models and all the analyses carried out with Proc Mixed of SAS in our examples can also be carried out with *lme* from S-Plus. The parms-statement used by SAS to fix the within trial variances corresponds with 'varFixed' in S-plus[55].
Several Stata macros have been written which implement some of the discussed methods[56, 57]. The Stata program *meta* of Sharp and Sterne[56] performs a standard fixed and random effects meta-analysis without covariates. The Stata command *metareg* of Sharp[57] extends this to univariate meta-regression. We are not aware of Stata programs that are capable to fit bivariate meta-regression models, but of course one can do an univariate meta-regression on the log odds ratios instead of a bivariate meta-regression on the log odds of the two treatment arms. However, such an analysis does not give any information about the relationship between the (true) log odds of the two arms.
Mlwin or MLn appears to be one of the most flexible method to fit mixed-effect regression models[54]. Although we do not have experience with this package, we guess that most if not all of the discussed models can be fitted in it.
Finally, in the free available Bayesian analysis software package BUGS, one can also execute all approximate likelihood analyses that were presented in this chapter. If vague prior distributions are used, the results are very similar. With BUGS it is also possible to fit the models using the exact likelihood, based on the binomial distribution of the number of events in a treatment arm. The reader is referred to Arends et al.[10] for examples and the needed BUGS syntax.

7 Bayesian statistics in meta-analysis

As we mentioned in section 4, putting uninformative Bayesian priors on all individual nuisance parameters as done in Thompson et al.[8], Daniels and Hughes[51], Smith, Spiegelhalter and Thomas[58] and Sharp and Thompson[30] can lead to inconsistent results as the number of nuisance parameters grows with the number of studies[36].
This observation does not imply that we oppose Bayesian methods. First of all, there is a lot of Bayesian flavour to random effects meta-analysis. The mixing distribution can serve as a prior distribution in the analysis of the results of a new trial. However, the prior is estimated from the data and not obtained by educated subjective guesses, that is why random effects meta-analysis can be seen as an example of the empirical Bayes approach. For each study, the posterior distribution given the observed value can be used to obtain empirical Bayes corrections.
In this chapter we describe estimating the mixing distribution by maximum likelihood. The maximum likelihood method has two drawbacks. First, in complex problems maximising the likelihood might become far from easy and quite time-consuming. Second, the construction of confidence intervals with the correct coverage probabilities can become problematic. We proposed the profile likelihood approach in the simple setting of section 3. For more complex problems, the profile likelihood gets very hard to implement.
When the maximum likelihood approach gets out of control (very long computing times, non-convergence of the maximisation procedure), it can be very profitable to switch to a Bayesian approach with vague priors on the parameters of the model in combination with Markov Chain Monte Carlo (MCMC) methods[45] that circumvent integration by replacing it by simulation. If one wants to use the MCMC technique in this context, the prior should be set on all parameters of the hierarchical model. Such a model could be described as a Bayesian hierarchical or Bayesian empirical Bayes model. For examples of this approach, see Arends et al.[10]. The difference with the approach of Thompson et al. [8, 30] is then that they assume that the true baseline log odds are a random sample of a fully specified flat normal distribution (e.g. N(0,10)), while we assume that the true log odds are sampled from a $N(\theta, \sigma)$ distribution with θ and σ parameters to be estimated, putting vague priors on them. So Thompson et al.'s model is a special case of our model. We prefer the parameters of the baseline risks distribution to be determined by the data.
For the examples discussed in this chapter, maximum likelihood was quite convenient in estimating the parameters of the model and getting a rough impression of their precision. It sufficed for the global analysis described here. If the model is used to predict outcomes of new studies, as in the surrogate marker setting of Daniels

and Hughes[51], nominal coverage of the prediction intervals becomes important and approximate methods can be misleading. MCMC can be very convenient, because the prediction problem can easily be embedded in the MCMC computations. An alternative is bootstrapping as described in Gail et al.[52].

8 Conclusions

We have shown that the general linear mixed model using an approximate likelihood approach is a very useful and convenient framework to model meta-analysis data. It can be used for the simple meta-analysis up to complicated meta-analyses involving multivariate treatment effect measures and explanatory variables. Extension to multiple outcome variables and multiple treatment arms is very straightforward. Software in widely available statistical packages can be used.

References

1. Normand S-L. Meta-analysis: formulating, evaluating, combining and reporting. *Statistics in Medicine* 1999; **18**:321-359.
2. Van Houwelingen H, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2272-2284.
3. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. Random effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14:**396-411.
4. Colditz GA, Brewer FB, Berkey CS, Wilson EM, Burdick E, Fineberg HV, et al. Efficacy of BCG vaccine in the prevention of tuberculosis. *Journal of the American Medical Association* 1994; **271**:698-702.
5. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177-188.
6. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* 1992; **11**(16):2077-2082.
7. McIntosh MW. The population risk as an explanatory variable in research syntheses of clinical trials. *Statistics in Medicine* 1996; **15**:1713-1728.
8. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741-2758.
9. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2883-2900.
10. Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24):3497-3518.
11. Fleiss JL. Analysis of data from multiclinic trials. *Controlled Clinical Trials* 1986; **7**:267-275.
12. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases* 1985; **27**:335-371.
13. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of clinical trials. *Statistics in Medicine* 1991; **10**:1665-1677.
14. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicne* 1996; **15**:619-629.
15. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 1987; **9**:1-30.

16. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1998; **17**:2635-2650.
17. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**:873-890.
18. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine* 2001; **20**:825-840.
19. Verbeke G. Linear Mixed Models for Longitudinal Data. In: Verbeke G, Molenberghs G, editors. *Linear Mixed Models in Practice.* Springer-Verlag: New York, 1997; pp 63-153.
20. Roger JH, Kenward MG, *Repeated measures using proc mixed instead of proc glm, Proceedings of the First Annual South-East SAS Users Group conference*. SAS Institute: Cary NC, 1993; pp 199-208.
21. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Annals of Internal Medicine* 1987; **107**(2):224-233.
22. Armitage P, Berry G, *Statistical Methods in Medical Research*. Blackwell Scientific Publications: Oxford, 1978.
23. Davey Smith G, Song F, Sheldon TA. Cholesterol lowering and mortality: the importance of considering initial level of risk. *British Medical Journal* 1993; **306**(6889):1367-1373.
24. Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). *Statistics in Medicine* 1994; **13**(3):293-296.
25. Brand R. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). *Statistics in Medicine* 1994; **13**(3):293-296.
26. Hoes AW, Grobbee DE, Lubsen J. Does drug treatment improve survival? Reconciling the trials in mild-to-moderate hypertension. *Journal of Hypertension* 1995; **13**(7):805-811.
27. Egger M, Smith GD. Risks and benefits of treating mild hypertension: a misleading meta-analysis? [comment]. *Journal of Hypertension* 1995; **13**(7):813-815.
28. Senn SJ. Relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**:1550.
29. Cook RJ, Walter SD. A logistic model for trend in 2 x 2 x K tables with applications to meta-analyses. *Biometrics* 1997; **53**(1):352-357.
30. Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Statistics in Medicine* 2000; **19**:3251-3274.
31. Carroll RJ, Ruppert D, Stefanski LA. *Measurement error in nonlinear models*. Chapman & Hall: London, 1995.

32. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735-738.
33. Kendall MG, Stuart A. *The advanced theory of statistics. Volume II: Inference and relationship.* Griffin: London, 1973.
34. Verbeke G, Lesaffre E. The effect of misspecifying the random effects distribution in linear models for longitudinal data. *Computational Statistics and Data Analysis* 1997; **23**:541-556.
35. Bernsen RMD, Tasche MJA, Nagelkerke NJD. Some notes on baseline risk and heterogeneity in meta-analysis. *Statistics in Medicine* 1999; **18**(2):233-238.
36. Van Houwelingen HC, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis (letter). *Statistics in Medicine* 1999; **18**:107-113.
37. Berlin JA, Antman EM. Advantages and limitations of meta-analytic regressions of clinical trials data. *Online Journal of Current Clinical Trials* 1994; **134**.
38. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**(6965):1351-1355.
39. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693-2708.
40. Bashore TR, Osman A, Heffley EF. Mental slowing in elderly persons: a cognitive psychophysiological analysis. *Psychology & Aging* 1989; **4**(2):235-244.
41. Jones DR. Meta-analysis of observational epidemiological studies: a review. *Journal of the Royal Society of Medicine* 1992; **85**(3):165-168.
42. Greenland S. A critical look at some popular meta-analytic methods. *American Journal of Epidemiology* 1994; **140**(3):290-296.
43. Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* 1999; **18**:643-654.
44. Engel B. A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal* 1998; **40**:141-154.
45. Gilks WR, Richardson S, Spiegelhalter DJ, *Markov Chain Monte Carlo in practice.* Chapman & Hall: London, 1996.
46. Laird NM. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* 1978; **73**:805-811.
47. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 1977; **39**(1):1-38.
48. Böhning D, Schlattman P, Linsay B. Computer-assisted Analysis of Mixtures. *Biometrics* 1992; **48**:283-304.

49. Eilers PHC, Marx BD. Flexible smoothing using B-splines and penalized likelihood. *Statistical Science* 1996; **11**:89-121.
50. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537-2550.
51. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**:1965-1982.
52. Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**(3):231-246.
53. Antczak-Bouckoms A, Joshipura K, Burdick E, Tulloch JFC. Meta-analysis of surgical versus non-surgical method of treatment for periodontal disease. *Journal of clinical periodontology* 1993; **20**:259-268.
54. Sutton AJ, Lambert PC, Hellmich M, Abrams KR, Jones DR. Meta-analysis in practice: A critical review of available software. In: Berry DA, Stangl DK, editors. *Meta-Analysis in Medicine and Health Policy*. Marcel Dekker: New York, 2000.
55. Pinheiro JC, Bates DM. *Mixed-effects models in S and S-Plus.* Springer-Verlag: Berlin, 2000.
56. Sharp S, Sterne J. Meta-analysis. *Stata Technical Bulletin* 1997; **38**(16):9-14.
57. Sharp S. Meta-analysis regression. *Stata Technical Bulletin* 1998; **42**(23):16-22.
58. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**:2685-2699.

4

Combining multiple outcome measures in a meta-analysis: an application

Abstract

In meta-analysis of clinical trials published in the medical literature it is customary to restrict oneself to standard univariate fixed or random effects models. If multiple endpoints are present, each endpoint is analysed separately. A few articles are written in the statistical literature on multivariate methods for multiple outcome measures. However, these methods were not easy to apply in practice, because self-written programs had to be used, and the examples were only two-dimensional. In this chapter we consider a meta-analysis on the effect on stroke-free survival of surgery compared to conservative treatment in patients with increased risk of stroke. Three summary measures per trial are available: short-term post-operative morbidity/mortality in the surgical group, long-term event rate in the surgical group and the event rate in the conservative group. We analyse the three outcomes jointly with a general linear MIXED model, compare the results with the standard univariate approaches and discuss the many advantages of multivariate modelling. It turns out that the general linear MIXED model is a very convenient framework for multivariate meta-analysis. All analyses could be carried out in standard general linear MIXED model software.

1 Introduction

Meta-analysis of clinical trials aims to combine estimates of treatment effect across related studies. Usually no individual patient data are available and use is made of summary data extracted from published literature and reports. The data per trial are summarised by one or more outcome measure estimates along with their standard errors. In practice mostly the data are reduced to one outcome measure per study, for instance the treatment effect estimated by means of an odds ratio. The data are then analysed by standard methods, using either a (univariate) fixed effect or, as preferred by most statisticians, a (univariate) random effects model[1]. If the summary data are multi-dimensional, then the data analysis is usually restricted to a number of separate univariate analyses. Raudenbush et al.[2] showed how to analyse two or more outcome measures jointly in a fixed effects multivariate linear model. Dear[3] used essentially the same method for combining survival curves in a meta-analysis, where each curve was characterised by estimated survival probabilities at two or more follow-up times. Van Houwelingen et al.[4] were the first to consider multivariate random effects meta-analysis. They introduced a bivariate linear random effects model for the joint analysis of one estimated outcome measure per treatment group. Essentially the same model was used by McIntosh[5] and Arends et al.[6] in the context of investigating the underlying risk as a source of heterogeneity in treatment effects across trials. Berkey et al.[7] introduced the general linear MIXED model as a general random effects regression method for meta-analysis of multiple outcomes. In a recent tutorial on advanced methods in meta-analysis[8], we adopted the general linear MIXED model as a general framework for multivariate meta-analysis and meta-regression. In fact this approach can be considered as a direct generalisation of the standard (univariate) DerSimonian-Laird[1] model to higher dimensions. To apply the model the estimated vector of outcome measures along with the corresponding estimated covariance matrix per trial is needed. The parameters are estimated with (restricted) maximum likelihood, acting as if the within trial covariance matrices are known. In this chapter we follow this approach in a meta-analysis about the effect on stroke-free survival of surgery versus conservative treatment in patients with high risk for stroke. Different from van Houwelingen et al.[8] and Berkey et al.[7], who had two outcome measures, we have three outcome measures: the event rate in the conservative treatment group, and the short-term and long-term event rate in the surgery group. The complication is that, because of the peri-operative mortality, the short-term stroke-free survival in the surgery group is lower than in the conservative group, while stroke-free survival on the long-term is in favour of the surgical treatment because of a lower event rate once the operation is survived. We use a

trivariate random effects model for the analysis of the data, and compare the results with univariate analyses. We show that the multivariate analysis is potentially much more informative than univariate analyses and can be carried out relatively easy in practice in standard software. Almost all models were fitted using Proc MIXED of SAS[9], while a few exact analyses were done with Proc NLMIXED. In this chapter we focus on the application. More theoretical details and background can be found in the recent tutorial on advanced methods in meta-analysis[8].
In section 2 we describe the data. In section 3 the models are introduced and the advantages of multivariate modelling are discussed. In section 4 we give the results, and the chapter ends with a discussion in section 5.

2 Data

In this chapter we analyse data from a meta-analysis of Vokó et al.[10] about the effect of carotid endarterectomy on all-cause mortality and stroke-free survival based on the aggregated data from 19 randomised trials. The vascular surgical procedure called carotid endarterectomy aims to remove the atherosclerotic plaque of the internal carotid artery and to restore the lumen of the vessel. To prevent cerebral infarction or death at people with increased levels of stenosis, one frequently performs a carotid endarterectomy[11, 12]. Although the operation mortality and morbidity is not negligible, the hope is that patients on average are better off because of lower event rates once the operation is survived. Several clinical trials comparing carotid endarterectomy plus best medical care with medical treatment alone have been done or are under way. Part of the studies that are published by now, are combined in the meta-analysis of Vokó et al.[10].
All selected trials were randomised clinical trials in which the indication of carotid endarterectomy was stroke prevention rather than treatment of acute stroke and in which the methodology was judged appropriate (no excessive loss to follow-up, symmetrical outcome assessment, analysed by treatment assignment from the moment of randomisation onwards). For further details about the selection of the trials we refer to Vokó et al.[10]. Together the 19 randomised clinical trials comprise in total 8991 patients being at increased risk of stroke, 4780 allocated to surgery and 4211 to conservative treatment. In this chapter we only look at stroke-free survival, so the event of interest is defined as stroke or death.
The basic data available for the 19 trials (i=1, ..., 19) were:

1. Number of patients in the surgical group (k_i) and number of events in the first month after operation (x_i) .

2. Number of events (y_i) and person years of follow-up (n_i) in the surgical group from 1 month post operation onwards.
3. Number of events (z_i) and person years of follow-up (m_i) in the conservative treatment group .

An implicit assumption was that the hazard rate is constant in the surgical group after the first month, and in the conservative group during the whole follow-up. This assumption could be somewhat relaxed, by splitting the time period in more periods and assuming a piecewise constant hazard rate. Of course, this results in more parameters to be estimated, i.e. one for each time interval, but the methods of this chapter remain applicable. In our case there were no data on events and person years on sub time intervals available.

The true event probability in the first month (called 'post-surgical risk' in the sequel) in the surgical group of trial i is denoted by π_i, estimated by the observed event probability $\hat{\pi}_i = x_i / k_i$. The true event rate after one month (called 'surgical long-term event rate' in the sequel) in the surgical group is denoted by λ_i. It is estimated by $\hat{\lambda}_i = y_i / n_i$. The true event rate (again called 'conservative long-term event rate' in the sequel) in the conservative treatment group is denoted by μ_i, estimated by $\hat{\mu}_i = z_i / m_i$. The data are given in Table 1.

Main questions were to compare the event-free survival of the two treatments depending on the length of the follow-up period and to investigate how the difference is modified by the level of underlying risk in the population. Secondary questions concerned the mean post-operative risk and the heterogeneity in it between trials, the difference between treatments in long-term event rate, and again how these are affected by the underlying risk.

3 Methods

3.1 Parameter transformations

As usual we transform the parameters such that the transformed parameters range from minus to plus infinity. This is more natural when random effects are employed. Moreover, the transformed parameters have better statistical properties if Wald type confidence intervals and tests are used.

The post-surgical risk parameter π_i is transformed to the log odds scale: $\omega_i = \ln(\pi_i/(1-\pi_i))$. The estimated log odds is denoted by $\hat{\omega}_i$. Its variance is estimated by

$$\text{var}(\hat{\omega}_i) = \frac{1}{x_i} + \frac{1}{k_i - x_i} \qquad (1)$$

Table 1. Data of the 19 clinical trials

Trial	Surgical group						Conservative group		
	First month			After one month					
	Events	Patients	Risk	Events	Person years	Event rate	Events	Person years	Event rate
	x_i	k_i	$\hat{\pi}_i$	y_i	n_i	$\hat{\lambda}_i$	z_i	m_i	$\hat{\mu}_i$
1	19	169	.112	26	564.56	.046	38	507.06	.075
2	7	20	.350	5	28.55	.175	7	50.60	.138
3	5	91	.055	5	79.01	.063	9	96.33	.093
4	5	78	.064	23	446.71	.051	16	366.17	.044
5	3	162	.019	57	920.04	.062	39	686.63	.057
6	14	200	.070	64	1141.92	.056	46	829.85	.055
7	18	190	.095	47	1103.46	.043	32	720.13	.044
8	22	350	.063	104	2005.75	.052	86	1403.83	.061
9	22	232	.095	60	1322.38	.045	48	790.53	.061
10	21	231	.091	68	1316.46	.052	69	985.82	.070
11	12	251	.048	86	1446.63	.059	71	900.71	.079
12	5	113	.044	38	650.83	.058	35	348.37	.100
13	45	678	.066	163	3065.42	.053	209	3120.17	.067
14	28	430	.066	92	1974.13	.047	156	1889.95	.083
15	19	328	.058	28	526.23	.053	80	489.87	.163
16	7	206	.034	49	529.50	.093	57	526.50	.108
17	0	15	.033	3	41.37	.073	1	44.33	.023
18	9	211	.043	41	513.16	.080	59	597.87	.099
19	22	825	.027	106	2004.36	.053	146	2076.45	070

(One trial, number 17, had zero events. As is usually done we added ½ , so $x_{17} = 0.5$.)
The long term event rates for the surgical and conservative treatment, respectively, are logarithmically transformed:

$\beta_i = \ln(\lambda_i)$, estimated by $\hat{\beta}_i = \ln(\hat{\lambda}_i)$ with estimated variance $\mathrm{var}(\hat{\beta}_i) = 1/y_i$ (2)

$\alpha_i = \ln(\mu_i)$, estimated by $\hat{\alpha}_i = \ln(\hat{\mu}_i)$ with estimated variance $\mathrm{var}(\hat{\alpha}_i) = 1/z_i$ (3)

The variances follow under the assumption of an exponential survival time distribution or constant hazard rate. The estimated transformed outcome measures are given in Table 2.

Table 2. Transformed data of the 19 trials

Trial	Surgical group				Conservative group	
	First month		After one month			
	Log odds post surgical risk		log long-term event rate		log long-term event rate	
	$\hat{\omega}_i$	$\text{var}(\hat{\omega}_i)$	$\hat{\beta}_i$	$\text{var}(\hat{\beta}_i)$	$\hat{\alpha}_i$	$\text{var}(\hat{\alpha}_i)$
1	-2.066	0.059	-3.078	0.038	-2.591	0.026
2	-0.619	0.220	-1.742	0.200	-1.978	0.143
3	-2.845	0.212	-2.760	0.200	-2.371	0.111
4	-2.681	0.214	-2.966	0.043	-3.130	0.063
5	-3.970	0.340	-2.781	0.018	-2.868	0.026
6	-2.587	0.077	-2.882	0.016	-2.893	0.022
7	-2.257	0.061	-3.156	0.021	-3.114	0.031
8	-2.702	0.490	-2.959	0.010	-2.793	0.012
9	-2.256	0.050	-3.093	0.017	-2.802	0.021
10	-2.303	0.052	-2.963	0.015	-2.659	0.014
11	-2.992	0.088	-2.823	0.012	-2.541	0.014
12	-3.073	0.209	-2.841	0.026	-2.298	0.029
13	-2.644	0.024	-2.934	0.006	-2.703	0.005
14	-2.664	0.038	-3.066	0.011	-2.494	0.006
15	-2.789	0.056	-2.934	0.036	-1.812	0.013
16	-3.347	0.148	-2.380	0.020	-2.223	0.018
17	-3.367	2.069	-2.624	0.333	-3.792	1.000
18	-3.111	0.116	-2.527	0.024	-2.316	0.017
19	-3.597	0.047	-2.940	0.009	-2.655	0.007

3.2 Univariate analyses

Surgical risk

To describe the post surgical risk among studies we adopt the standard (univariate) random effects model of DerSimonian and Laird[1].

$$\hat{\omega}_i \cong \text{N}(\omega_i, \text{var}(\hat{\omega}_i)) \qquad (4)$$

$$\omega_i \cong \text{N}(\omega, \sigma_\omega^2)$$

We will refer to the two submodels as the measurement error model and the structural model, respectively. Here ω_i is the true logit(post-surgical risk) for trial *i*. The ω_i's may vary over trials, and are assumed to follow a normal distribution with mean ω and standard deviation σ_ω, the latter characterising the heterogeneity among trials. The ω_i's are not observed, but estimated by $\hat{\omega}_i$, which are assumed to have a

normal distribution with mean ω_i and variance given by (1). The normality assumption for $\hat{\omega}_i$ is usually justified by large enough sample sizes. In fact, it is only assumed that the likelihood of $\hat{\omega}_i$ is well approximated by a normal distribution likelihood, which is a somewhat weaker assumption. The normality assumption for the true ω_i's, although standard, is more crucial, but for a larger number of trials the inference on ω and σ_ω is robust against misspecification of this distribution[13, 14]. The parameters ω and σ_ω^2 are estimated by standard maximum likelihood or restricted maximum likelihood methods[14], doing as if the study specific variances are known.

The model can be fitted using any standard linear MIXED model program provided that it is possible to fix the residual variances at user specified values. We used the procedure Proc MIXED of the SAS package[9].

In the recent advanced meta-analysis tutorial[8] exact approaches were discussed that can be fitted in very special and relatively simple cases. Where feasible we will do that, in order to compare the results with the approximate approach. When in model (4) the approximate measurement error model is replaced by the exact one we get:

$$x_i \cong \text{Bin}(k_i, \frac{\exp(\omega_i)}{1+\exp(\omega_i)}) \qquad (5)$$
$$\omega_i \cong \text{N}(\omega, \sigma_\omega^2)$$

This is a logistic-normal random effects model which could be fitted for instance with EGRET[15] or MIXOR[16, 17]. We fitted the model using Proc NLMIXED of SAS[9]. (Since x_i is allowed to be zero in this model, we changed x_{17}=0.5 back to x_{17}=0.)

Long term risks

To compare the long term risks between the treatments we look at the difference $\delta_i = \alpha_i - \beta_i$ and assume again the standard random effects model[1]:

$$\hat{\delta}_i \cong \text{N}(\delta_i, \text{var}(\hat{\delta}_i)) \quad \text{with} \quad \text{var}(\hat{\delta}_i) = \text{var}(\hat{\alpha}_i) + \text{var}(\hat{\beta}_i) \qquad (6)$$
$$\delta_i \cong \text{N}(\delta, \sigma_\delta^2)$$

The estimated variances are computed with formula (2) and (3). The parameters δ and σ_δ^2 are estimated by maximum likelihood, assuming the study specific variances to be known.

In this case it is also possible to fit the exact measurement error model. Exploiting the fact that the conditional distribution of z_i given z_i+y_i is binomial with parameters z_i+y_i and $\mu_i m_i / (\mu_i m_i + \lambda_i n_i)$, respectively, the model can be written as:

$$z_i \cong \text{Bin}(z_i + y_i, \frac{\exp(\log(m_i/n_i) + \delta_i)}{1+\exp(\log(m_i/n_i) + \delta_i)}) \qquad (7)$$

$$\delta_i \cong \mathrm{N}(\delta, \sigma_\delta^2),$$

This is a logistic-normal random effects model with $\log(m_i / n_i)$ as offset variable. Again we fitted this model with Proc NLMIXED.

Cumulative survival ratio

To compare event free survival probabilities between the treatments over a fixed follow-up interval (0, *t*) we look at the ratio of the cumulative *t*-year survival probabilities:

$$\mathrm{CSR}_i(t) = \frac{(1-\pi_i)\exp(-t'\lambda_i)}{\exp(-t\mu_i)}$$

where *t'* = *t*-1/12 and i the number of the trial. The choice of *t* is arbitrary, but in the original meta-analysis of Vokó et al.[10] focus was on *t* = 3 years since most of the trials had a mean follow-up duration of about 3 years.

For the analysis it is natural to work with the logarithmically transformed parameter:

$$\rho_i(t) = -\log(1+\exp(\omega_i)) - t'\exp(\beta_i) + t\exp(\alpha_i)$$

The ρ_i's are estimated by plugging in the estimates of ω_i, α_i, and β_i, while the variances are estimated using the delta-method by:

$$\mathrm{var}(\hat{\rho}_i) = \hat{\pi}_i^2\,\mathrm{var}(\hat{\omega}_i) + t'^2\,\hat{\lambda}_i^2\,\mathrm{var}(\hat{\beta}_i) + t^2\,\hat{\mu}_i^2\,\mathrm{var}(\hat{\alpha}_i) \qquad (8)$$

Again we adopt the standard random effects model for the log(cumulative *t*-years survival ratio):

$$\begin{aligned} \hat{\rho}_i(t) &\cong \mathrm{N}(\rho_i(t), \mathrm{var}(\hat{\rho}_i(t))) \\ \rho_i(t) &\cong \mathrm{N}(\rho(t), \sigma_\rho^2(t)) \end{aligned} \qquad (9)$$

For fixed value of *t*, the parameters $\rho(t)$ and $\sigma_\rho^2(t)$ are again estimated by maximum likelihood, assuming the study specific variances (8) to be known.

3.3 Multivariate analyses

In this subsection we introduce a multivariate model in which all three outcome measures are analysed simultaneously. The model is a direct generalisation of the above univariate random effects models. Again we work with the transformed parameters. Given the true trial specific outcome measures we assume that the estimates follow a multivariate normal distribution.

$$\begin{pmatrix} \hat{\omega}_i \\ \hat{\beta}_i \\ \hat{\alpha}_i \end{pmatrix} \cong \mathrm{N}\left[\begin{pmatrix} \omega_i \\ \beta_i \\ \alpha_i \end{pmatrix}, \begin{pmatrix} \mathrm{var}(\hat{\omega}_i) & 0 & 0 \\ 0 & \mathrm{var}(\hat{\beta}_i) & 0 \\ 0 & 0 & \mathrm{var}(\hat{\alpha}_i) \end{pmatrix} \right] \qquad (10.1)$$

In general the covariances might be non-zero, but in our application the correlations are zero because the likelihood factorises in three parts each involving only one parameter.

For the true outcome measures we assume a multivariate distribution as well.

$$\begin{pmatrix} \omega_i \\ \beta_i \\ \alpha_i \end{pmatrix} \cong \mathrm{N}\left[\begin{pmatrix} \omega \\ \beta \\ \alpha \end{pmatrix}, \begin{pmatrix} \sigma_{\omega\omega} & \sigma_{\omega\beta} & \sigma_{\omega\alpha} \\ \sigma_{\omega\beta} & \sigma_{\beta\beta} & \sigma_{\alpha\beta} \\ \sigma_{\omega\alpha} & \sigma_{\alpha\beta} & \sigma_{\alpha\alpha} \end{pmatrix} \right] \tag{10.2}$$

Marginally this model assumes just a standard DerSimonian-Laird model for each outcome measure. The parameters can be estimated by standard likelihood, doing as if the variances of the trial specific outcome measures are known. We again used SAS Proc MIXED to fit the model.

The multivariate modelling has several advantages. First, instead of doing a number of univariate analyses each tailored to one specific question, the multivariate approach gives a complete and concise description of all data at one stroke. Once the model is fitted, it immediately gives the estimated mean post-surgical risk, the long-term risks under both treatments and the between trial variances of these parameters. Inference on derived parameters can readily be carried out. For instance, the estimate of the difference in log(long-term risk) can easily be computed, and the associated *P*-value and confidence interval follow directly from the covariance matrix of the estimates. In the univariate approach, a separate analysis had to be done for estimating the log(cumulative survival ratio) over (0,*t*) for each value of *t* that was of interest. The multivariate approach yields an estimate and confidence interval for the typical log(cumulative survival ratio) over (0,*t*) as a relatively simple function of *t*. Second, the multivariate approach yields the estimated correlations between the outcome measures. This can lead to more insight. For instance it might be interesting to know whether high post-surgical risks are associated with higher or rather with lower long-term risks, either under the conservative or under the surgical treatment. One would probably want to adjust the latter association for the long term risk under the conservative treatment, then one looks at the partial correlation between ω_i and α_i given β_i. The third advantage that we mention is related to the previous point. Often one is interested in whether a measure of treatment benefit is modified by some measure of baseline risk. For instance, is the difference in log(long term risk)'s between treatments associated with larger baseline risks as measured by the long term risk under the conservative treatment? Or, is the log(cumulative survival ratio) over (0,*t*) modified by the baseline risks as measured by the long term risk under the conservative treatment? For this type of questions the univariate analyses of the previous section fall short and multivariate modelling is necessary. We elaborate on this in section 4.2. The fourth advantage of the multivariate approach that we mention

is that it is capable to deal adequately with incomplete trials, that is when one or two outcome measures are missing, and therefore makes more efficient use of the data. If the missing outcomes are missing at random but not completely at random, the multivariate approach might also be more valid than the univariate analyses that necessarily leave out the incomplete trials. In our application the event of interest was defined stroke or death. Most trial reports, but not all, also give the outcome event death alone. Probably both outcomes will be highly correlated. Therefore, if one is interested in the effect of treatment on the endpoint death, it would be advantageous to carry out a multivariate analysis with both the outcome death and outcome stroke or death. One can also think of a situation where one has a surrogate endpoint for all trials, and a smaller number of trials reporting both the 'true' endpoint as well as the surrogate. Then one is specifically interested in the trial-level correlation between both endpoints[18, 19].

4 Results

4.1 Univariate analyses

The univariate models (4), (6) and (9) were fitted using Proc MIXED of SAS[9]. Since this is not completely trivial, we refer the reader to the recent tutorial on advanced meta-analysis methods[8] for explanation on how to do that.

The estimated mean log odds of a post-surgical event was $\hat{\omega}$ = -2.681, with standard error 0.133. So on the original scale the estimated mean post-surgical risk is 0.064 with approximate 95% confidence interval (0.050, 0.082). The estimated between trials variance on the log odds scale was estimated as 0.224, giving an approximate 95% coverage interval of the true post-surgical risks of (0.026, 0.148). This indicates quite a large between trial variation in post-surgical risks. The likelihood ratio test on H_0: σ_ω^2 = 0 was borderline statistically significant (P = 0.08). Proc MIXED also gives a Satterthwaite approximation based 95% confidence interval for the between trials variance, (0.097, 0.950).

The exact measurement error model fitted by SAS Proc NLMIXED gives an estimated mean log odds of a post-surgical event $\hat{\omega}$=-2.739, with standard error 0.130. This corresponds on the original scale to an estimated mean post-surgical risk of 0.065 with a 95% confidence interval (0.049, 0.085). The estimated between trials variance on the log odds scale was 0.211, resulting in an 95% coverage interval of the true post-surgical risks of (0.026, 0.159). All of these estimates are quite similar to the approximate likelihood estimates.

The estimate of the mean difference in log(long-term event rate)'s was $\hat{\delta}$ = 0.277 with standard error 0.061, so on average the long-term event rate of the surgical treatment was highly significantly better than of the conservative treatment. The estimated hazard ratio is 1.32 with approximate 95% confidence interval (1.17, 1.49). The estimated between trials variance in true log(long-term event rate difference)'s is 0.0268 (95% confidence interval (0.0078, 0.6425)), significantly different from zero at the 5% level (P = 0.04). The approximate 95% coverage interval of the true long-term event rate ratios is (0.96, 1.82), again indicating quite a large between trials variation.
The exact measurement error model fitted by SAS Proc NLMIXED gives an estimated mean difference in log(long term event rate)'s equal to $\hat{\delta}$ =0.278, with standard error 0.064. The estimated hazard ratio is 1.32 with 95% confidence interval (1.15, 1.51). This is very similar to the approximate likelihood estimates. The estimated between trials variance in the true log(long term event rate difference)'s is 0.032, significantly different from zero (P = 0.02, likelihood ratio test). The 95% coverage interval of the true long-term event rate ratios is (0.93, 1.58). This is all very similar to the results based on the approximate likelihood.
Model (9) was fitted for a number of different values of *t*. The results for some selected values of *t* are given in Table 3.

Table 3. Results of the univariate random effects model for the cumulative *t*-years survival probability ratio for some selected values of *t*.

	***t* (years)**		
	1	3	8
Estimated mean log CSR $\hat{\rho}(t)$	-0.0379	-0.00756	0.0747
Standard error of $\hat{\rho}(t)$	0.00844	0.0155	0.0356
P-value for H_0: $\rho(t)=0$	<0.0001	0.63	0.036
Between trials variance $\hat{\sigma}_\rho^2(t)$	0.000744	0.00232	0.0111
LR test p-value for H_0: $\sigma_\rho^2 = 0$	0.0002	0.008	0.026
Estimated cumulative survival ratio CSR(*t*)	0.963	0.9925	1.078
95% confidence interval for CSR(*t*)	0.947, 0.979	0.963, 1.023	1.005, 1.155

In the previous analysis it was seen that the long-term event rate was better for the surgical treatment. However, for relative short follow-up times, the event free cumulative survival probability is in favour of the conservative treatment because of the post surgical risk. For example, from Table 3 it is seen that for one year follow-up duration survival is very significantly worse for the surgical treatment. For longer

follow-up duration the balance is in favour of the surgical treatment. From Table 3 it is seen that for t = 3 years follow-up the estimated event free survival probability is about equal for both treatments. From about t = 8 years cumulative survival for the surgical treatment is significantly better than for the conservative treatment. In Figure 1 the estimated cumulative survival ratio and its 95% confidence interval is given as a function of t. Moreover the approximate 95% coverage interval is given, i.e. the interval in which the true cumulative survival probability ratio of a new trial will lie with about 95% probability. It is seen that the estimated length of follow-up for which the two treatments are equivalent is 3.5 years with a 95% confidence interval running from 2.0 to 7.4 years. These values were determined by using a fine grid of values of t and running the analysis for each value of t.

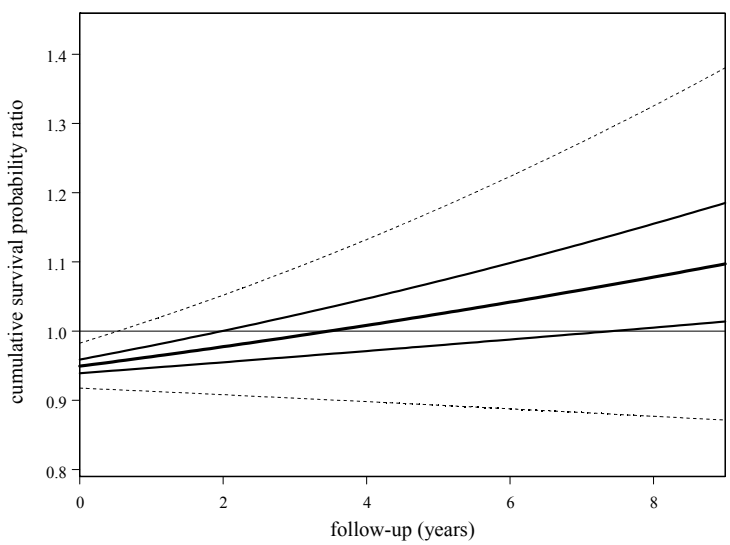


Figure 1. Estimated mean cumulative survival probability ratio, based on *univariate* analyses for different lengths of follow-up (bold curve). The inner two curves give the 95% confidence interval, and the outer curves approximate 95% coverage intervals.

4.2 Multivariate analysis

The multivariate model could be fitted using SAS Proc MIXED as well. The main results of the multivariate meta-analysis are given in Table 4.

Table 4. Some results of the trivariate model (10.1) and (10.2).

	Outcome measure		
Estimate	Logit of post-surgical risk (ω)	Log long-term event rate of surgical treatment (β)	Log long-term event rate of conservative treatment (α)
Mean	-2.707	-2.891	-2.573
Standard error	0.1337	0.0440	0.0777
Between trials variance	0.2299	0.0167	0.0852

The full estimated covariance matrix of the estimated mean outcome measures is:

$$\operatorname{cov ar}\begin{pmatrix}\hat{\omega}\\ \hat{\beta}\\ \hat{\alpha}\end{pmatrix}=\begin{pmatrix}0.017888 & -0.002154 & -0.0005938\\ -0.002154 & 0.001939 & 0.001925\\ -0.0005938 & 0.001925 & 0.006036\end{pmatrix} \quad (11)$$

The estimated covariance matrix of the random effects is:

$$\operatorname{cov ar}\begin{pmatrix}\omega_i\\ \beta_i\\ \alpha_i\end{pmatrix}=\begin{pmatrix}0.2299 & -0.03666 & -0.01144\\ -0.03666 & 0.01675 & 0.03220\\ -0.01144 & 0.03220 & 0.08519\end{pmatrix} \quad (12)$$

This covariance matrix turned out to be positive semi-definite. The estimated correlation matrix of the random effects is:

$$\operatorname{corr}\begin{pmatrix}\omega_i\\ \beta_i\\ \alpha_i\end{pmatrix}=\begin{pmatrix}1 & -0.59 & -0.08\\ -0.59 & 1 & 0.85\\ -0.08 & 0.85 & 1\end{pmatrix}$$

SAS Proc MIXED also gives the estimated covariance matrix of the estimates of the random effects parameters (not shown).

Notice that for the logit of the post-surgical risk the result is almost identical to the above given univariate analysis. This is also true for the other two outcomes (univariate results not shown).

A number of questions could be answered using the results of the multivariate analysis. Let us start with comparing the two long-term event rates. The estimated mean difference in log(long-term event rate)'s is 0.318 with standard error $\sqrt{0.001939+0.006036-2\cdot 0.001925}=0.064$, not much different from the univariate analysis. The associated between trials variance is estimated as 0.01675+0.08519 – 2·0.0322 = 0.0375, slightly larger than from the univariate analysis.

We now look at the *t*-years cumulative survival ratio. In the univariate approach we had to repeat the analysis for each value of *t* of interest. An advantage of the multivariate model is that the estimated *t*-years cumulative survival ratio can be given as a simple function of *t*:

$\hat{CSR}(t) = \exp(-0.06 + 0.021t) = 0.94\,e^{0.021 \cdot t}$

Note that the interpretation of this is a little bit different from the CSR(*t*) from the univariate analysis. In the univariate approach the mean log(CSR(*t*)) was estimated. After exponentiating it can be interpreted as the estimated CSR(*t*) for a trial with average log(CSR(*t*)), or as the estimated median CSR(*t*). Now we have estimated the CSR(*t*) for the 'typical' clinical trial having average post-surgical log odds and average log(long-term risk) under both treatments. Of course, based on the multivariate analysis it would be possible to compute the analogue of the univariate parameter estimate, by integrating the estimated log(CSR(*t*)) over the estimated trivariate normal distribution of the random effects, but that is not very simple and there is no need to that since the present parameter estimate is perfectly interpretable. The value of *t* for which the cumulative survival probability over (0,*t*) is equal for both treatments is estimated as 2.88, somewhat smaller than found in the univariate analyses.

The estimated variance of its logarithm is computed from the analogue of (8) and the covariance matrix (11), and is a simple quadratic function of *t*:

$\mathrm{Var}(\ln(\mathrm{CSR}(t))) = (0.712 - 0.0893\,t + 0.248\,t^2)/10^4$

The approximate 95% confidence interval for CRS(*t*) is thus given by:

$$0.94\exp(0.021\,t \pm 1.96\sqrt{(0.712 - 0.0893t + 0.248t^2)/100}) \qquad (13)$$

The confidence interval for the value of *t* for which the event free survival probability is equal for both treatments is conveniently computed by converting (13) and turns out to be (1.83, 5.50). As an illustration we give in Figure 2 the estimated CSR(*t*) and its corresponding 95% confidence interval.

An advantage of the above multivariate analysis is that the correlations between the different outcome measures are estimated. Notice that it would not be adequate to look at the simple correlations between the observed outcome measure estimates, since we are interested in the correlations between the underlying true trial specific outcomes. Moreover, the observed outcome measures have different precisions between trials and the errors might in general be correlated too (although this was not the case in our application), so that within and between trial correlation would be mixed up.

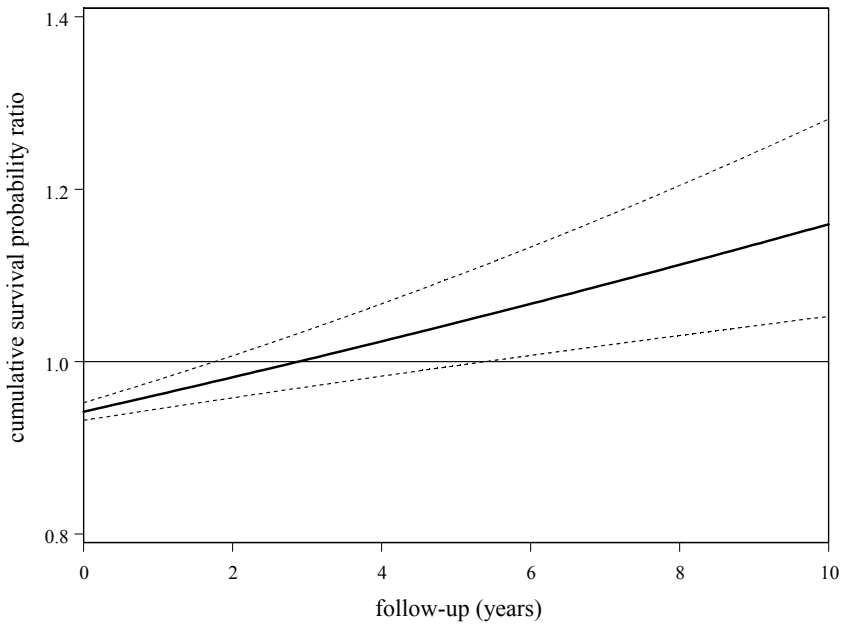


Figure 2. Estimated cumulative survival probability ratio, based on the *multivariate* analysis as a function of length of follow-up. The dotted curves give the 95% confidence band.

An interesting finding in our example is that there was almost no correlation between the post-surgical risk and the long-term event rate under conservative treatment. This is an indication that the post-operative risk is not higher in high risk populations, in contrast to what was expected beforehand. Another finding is that there is a moderately high negative correlation of –0.59 between the post-surgical risk and the long-term risk under the surgical treatment. This is an indication that the most vulnerable patients tend to have an event in the first month after surgery leading to a selected group of patients with good long-term prognosis, a kind of 'survival of the fittest' phenomenon. Probably one would want to adjust this correlation for the event rate in the conservative treatment. The partial correlation turns out to be equal to –1, which is an even stronger indication of this selection phenomenon that the patients with a post operative event are probably the ones that otherwise would have had an event later on.

Above it appeared that there is quite some variation among trials in treatment effect measures such as the difference in log(long-term events rate)'s and the *t*-years cumulative survival probability ratios.

One possibility of exploring this heterogeneity would be to make use of trial level covariates. The above multivariate model is straightforwardly extended with covariates, which might be different for different outcomes. In the application of this chapter there were no covariates available. In the absence of covariates, although not only then, it is quite common to consider whether there is any association between patients' underlying risk of the event in question and the treatment effect measure. The underlying risk is a convenient and clinically relevant trial-level measure which can be interpreted as a summary of a number of unmeasured patient characteristics. In our application the log(long-term event rate) α_i is the straightforward choice for the baseline risk measure. Simply regressing the estimated treatment on the observed baseline risk measure would be a mistake for several reasons (see for instance Sharp[20]), and a more sophisticated approach is needed. A number of articles has been written on how to estimate the relation between treatment effects and underlying risk in meta-analyses[5, 6, 21-25]. The approach of this chapter is in the spirit of Arends et al.[6] and McIntosh [5], and is easily carried out using the results of the multivariate analysis.

As a first example, suppose that one is interested in whether the long-term treatment effect is different between low and high risk populations, or, in other words, whether the long-term treatment effect depends on the long term event rate in the conservative treatment group. Then it is natural to look at the regression line of $\delta_i = \alpha_i - \beta_i$ on α_i, which is given by:

$$\delta_i = \delta + (1 - \frac{\sigma_{\alpha\beta}}{\sigma_{\alpha\alpha}})(\alpha_i - \alpha)$$

All ingredients that we need are available from the multivariate results above. The estimated regression line is δ_i = 1.918 + 0.622 α_i. The estimated standard error of the slope is 0.2924, computed with the delta method using the covariance matrix of the estimated covariance matrix (not shown). We conclude that the slope differs significantly from zero, so the long term event risk ratio declines with increasing baseline event rate in favour of the surgical treatment. This is illustrated in Figure 3. A confidence band for the regression line might be computed with the delta-method, using the estimated covariance matrices of the fixed effects and covariance parameters. The residual standard deviation is 0.069 and the percentage explained variance is quite high, 87.5%. The fit of the regression line to the observed long-term event ratios appears to be quite good, except maybe for the very small trial number 17 at the right hand side below.

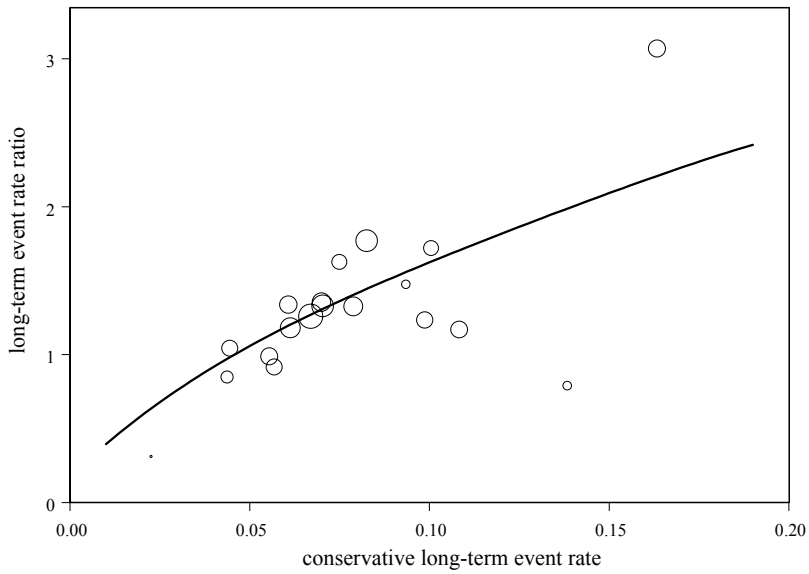


Figure 3. Observed long-term event rate ratios (conservative relative to surgical treatment) plotted against observed long-term event rate in the conservative treatment group and estimated regression line of true long-term event rate ratio on true conservative long-term event rate. Area of circles is proportional to the number of long-term events.

Another relationship of interest is between the *t*-years cumulative survival ratio and the underlying risk. Therefore we look at the conditional distribution of (ω_i, β_i) given α_i, which is bivariate normal with mean

$$\begin{pmatrix} \omega + \frac{\sigma_{\omega\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i - \alpha) \\ \beta + \frac{\sigma_{\beta\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i - \alpha) \end{pmatrix}$$

The typical log(CRS(*t*)) therefore is

$$\rho(t;\alpha_i) = -\ln\left(1+\exp(\omega + \frac{\sigma_{\omega\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i - \alpha))\right) - (t - \frac{1}{12})\exp\left(\beta + \frac{\sigma_{\beta\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i - \alpha)\right) + t\exp(\alpha_i)$$

Using the results of the multivariate analysis this is estimated by:

$$\hat{\rho}_i(t;\alpha_i) = -\ln\left(1+\exp(-3.053 - 0.134 \quad \alpha_i)\right) - (t - \frac{1}{12})\exp\left(-1.918 + 0.378 \quad \alpha_i\right) + t\exp(\alpha_i)$$

Given α_i the estimated break-even value of t, that is the value of t for which the survival probability over $(0, t)$ is equal for both treatments, is given by

$$t_{break-even} = \frac{\ln\left(1+\exp(\omega+\frac{\sigma_{\omega\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i-\alpha))\right)-\exp\left(\beta+\frac{\sigma_{\beta\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i-\alpha)\right)/12}{\exp(\alpha_i)-\exp\left(\beta+\frac{\sigma_{\beta\alpha}}{\sigma_{\alpha\alpha}}(\alpha_i-\alpha)\right)}$$

This is estimated by

$$\hat{t}_{break-even} = \frac{\ln\left(1+\exp(-3.053-0.134\ \ \alpha_i)\right)-\exp\left(-1.919+0.378\ \ \alpha_i\right)/12}{\exp(\alpha_i)-\exp\left(-1.919+0.378\ \ \alpha_i\right)}$$

The estimated break-even point is positive as long as the long-term event rate for the surgical treatment is lower than the predicted long-term event rate for the conservative treatment. As an illustration we plotted in Figure 4 the observed and predicted break-even point t against the observed event rate in the conservative treatment, for the trials with positive observed and predicted break-even times. Confidence intervals might be computed with the delta-method, using the estimated covariance matrices of the fixed effects and covariance parameters. The fit of the observed break-even times to the predicted break-even times appears to be quit good.

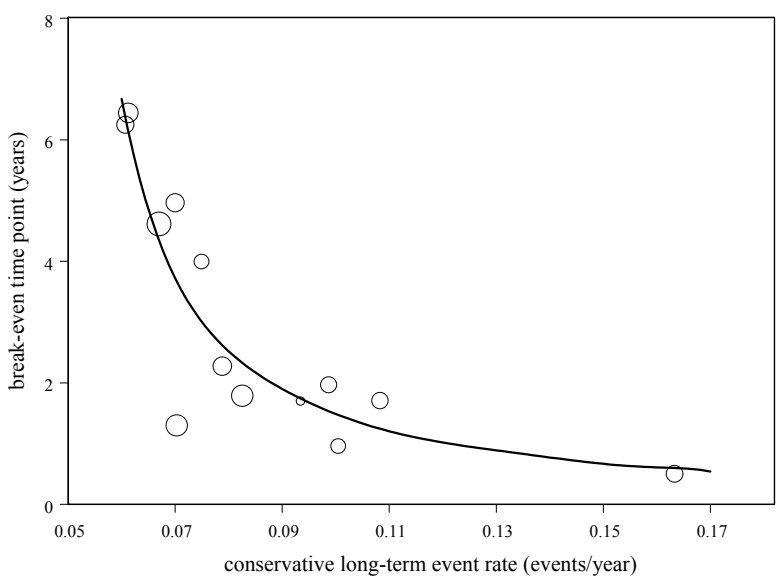


Figure 4. Observed and predicted break-even time against long-term event rate for the conservative treatment. Area of circles is proportional to the total number of events in a trial.

In Figure 5 the estimated CSR(t) is given as a function of the true conservative long-term event rate μ_i =exp(α_i) for selected values of t. From the picture it can be seen for instance that in a population with baseline incidence over about 8 events per 100 person-years the typical 3-years survival probability under surgical treatment is better than under conservative treatment. Again confidence intervals for $\rho(t;\alpha_i)$ can be constructed via the delta-method. Notice that the predicted 3-years survival probability ratio curve fits the observed 3-years survival probability ratios very well, except for one outlier, the very small trial number 17.

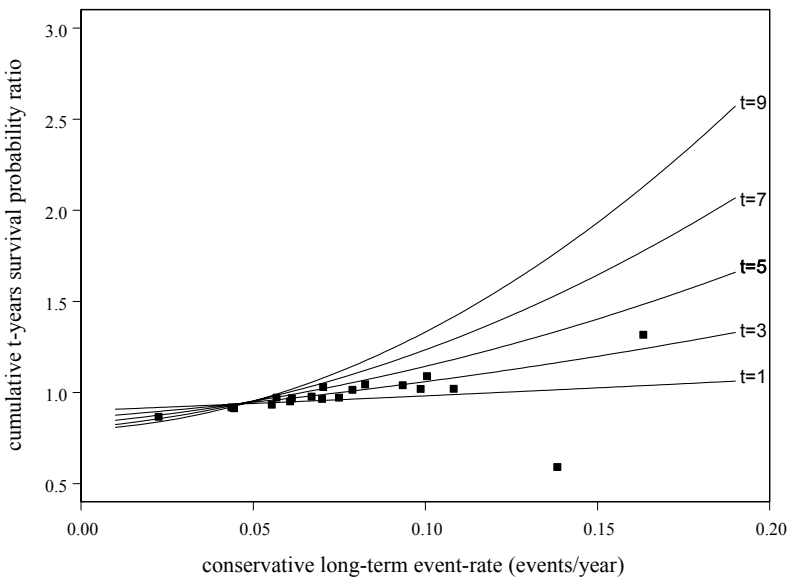


Figure 5. Estimated cumulative survival ratio probability ratio (surgical relative to conservative treatment) for different follow-up periods (0,t) (t in years). The points are the observed 3-years cumulative survival probability ratios plotted against the observed long-term event rate in the conservative treatment group.

5 Discussion

To our knowledge this is the first example of a multivariate random effects meta-analysis combining more than two outcomes. The model that we used is quite generally applicable. In our application we had no covariates available that could explain heterogeneity between trials. If they are available, they can be used without any further difficulties. In our example, the different outcome measures were

independent within trials. Also this is no limitation of the method and the presence of correlations can easily be accommodated. For a bivariate meta-analysis example where this is done we refer to van Houwelingen et al.[8] We demonstrated the advantages of the multivariate analysis upon the univariate analyses. One multivariate analysis yielded much more information than a number of separate univariate analyses. The multivariate analysis revealed the relations between the different outcomes, gave simple expressions for estimation of derived treatment effect parameters such as for instance the cumulative survival probability ratio as a function of follow-up duration. Furthermore, the results of the multivariate modelling enabled us to easily estimate the relation of different treatment effect parameters and the underlying risk. We did not have missing outcome measures in our example, but our method allows them. In other applications this can increase efficiency compared with the analysis restricted to only the trials with a complete set of outcome measures. This also makes the model very useful in modelling the relationship between surrogate and true endpoints in a meta-analysis with a mix of trials, some of them reporting both the surrogate and the true outcome and the others only the surrogate outcomes.

Fortunately, the multivariate model can easily be fitted in standard general linear MIXED model programs, although not every program will have the appropriate options. We used SAS Proc MIXED, but we guess that other packages such as S-Plus or MLWin might also be used, although we do not have extensive experience with these programs. The essential requirement is that the residual variances can be fixed at arbitrary values per individual trial[8].

We fitted the multivariate model using straightforward likelihood, but approximate because we acted as if the residual variances were estimated without error. In a few special univariate cases an exact likelihood was possible as well. In those cases the results turned out to be very similar. At present, an exact likelihood approach is not feasible in the multivariate case. An alternative approach would be to fit the model using Bayesian methods. This can for instance be done in the free available Bayesian analysis package BUGS[26]. One advantage is that the exact likelihood can be used by specifying the distribution for the outcome measure, in our example a binomial distribution for the number of post-operative events and a Poisson distribution for the events on long-term in both treatment groups. Another advantage is that, since BUGS uses MCMC methods to sample from the posterior distribution of all parameters, the inference based on the results of the fitted model can be easily build in. In section 4.2 we computed by hand derived results such as the cumulative survival probability ratio and the break-even point, and their regression with the underlying risk. The estimates were quite easily computed, but the standard errors and confidence intervals using the delta-method are more cumbersome, especially for the regression

relationships. In BUGS this kind of derived inference including the (Bayesian) confidence intervals can be done very conveniently in the program. Some results of the BUGS analysis are presented in Table 5.

Table 5. Some results of the trivariate model fitted by BUGS.

	Outcome measure		
Estimate	Logit of post-surgical risk (ω)	Log long-term event rate of surgical treatment (β)	Log long-term event rate of conservative treatment (α)
Mean	-2.760	-2.907	-2.604
Standard error	0.1322	0.0464	0.0805
Between trials covariance matrix of the random effects	$\text{covar}\begin{pmatrix}\omega_i\\ \beta_i\\ \alpha_i\end{pmatrix} = \begin{pmatrix} 0.2226 & -0.0356 & -0.0266 \\ -0.0356 & 0.0175 & 0.0224 \\ -0.0266 & 0.0224 & 0.0974 \end{pmatrix}$		

The results are very similar to those of the approximate likelihood approach (Table 3). As a further illustration we reproduce the analogues of Figures 2 to 5 now with the Bayesian approach, see Figure 6. In Figure 6 the dotted lines represent the approximate likelihood estimate, while the solid lines represent the results of the BUGS analysis. Again the results are very comparable. Of course, a practical disadvantage of this approach is that fitting this kind of models in a program like BUGS can be quite time consuming, and therefore the approach presented in this chapter is much more practical.

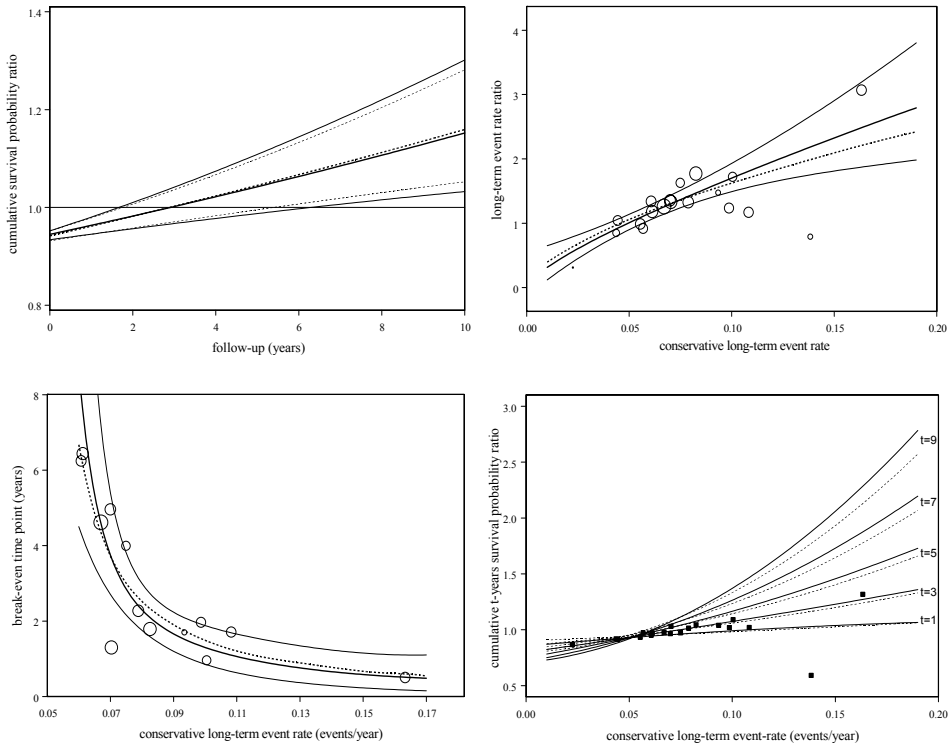


Figure 6. Analogues of Figures 2 to 5, with approximate likelihood as well as the Bayesian approach. The dotted lines represent the approximate likelihood estimates, the solid lines represent the results of the BUGS analysis.

References

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177-188.
2. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychological Bulletin* 1988; **103**(1):111-120.
3. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994; **50**(4):989-1002.
4. Van Houwelingen JC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2272-2284.
5. McIntosh MW. The population risk as an explanatory variable in research syntheses of clinical trials. *Statistics in Medicine* 1996; **15**:1713-1728.
6. Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24):3497-3518.
7. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537-2550.
8. Houwelingen HC van, Arends L, Stijnen T. Tutorial in Biostatistics: Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589-624.
9. SAS [program]. Version 8.0. Cary, N.C.: SAS Institute, Inc., 1999.
10. Vokó Z. *Etiology and prevention of stroke. The Rotterdam Study [Thesis].* Erasmus University Rotterdam: Rotterdam, 2000; pp 59-75.
11. Dyken ML, Pokras R. The performance of endarterectomy for disease of the extracranial arteries of the head. *Stroke* 1984; **15**:948-950.
12. Gillum RF. Epidemiology of carotid endarterectomy and cerebral arteriography in the United States. *Stroke* 1995; **26**:1724-1728.
13. Verbeke G, Lesaffre E. The effect of misspecifying the random effects distribution in linear models for longitudinal data. *Computational Statistics and Data Analysis* 1997; **23**:541-556.
14. Verbeke G. Linear Mixed Models for Longitudinal Data, In: Verbeke G, Molenberghs G, editors. *Linear Mixed Models in Practice*. Springer-Verlag: New York, 1997; pp 63-153.
15. EGRET for Windows [program]. Version 2.0.3. Cambridge: CYTEL Software Corporation, 1999.
16. MIXOR [program]. Version 2.0. Hedeker D, Gibbons RD. Chicago, 1996.

17. Hedeker D, Gibbons RD. MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* 1996; **49**:157-176.
18. Buyse M, Molenberghs G, Burzykowsky T, Renard D, Geys H. The validation of surrogate endpoints in met-analysis of randomized experiments. *Biostatistics* 2000; **1**(1):49-68.
19. Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**(3):231-246.
20. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735-738.
21. Brand R. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). *Statistics in Medicine* 1994; **13**(3):293-296.
22. Senn SJ. Relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**:1550.
23. Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Statistics in Medicine* 2000; **19**:3251-3274.
24. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741-2758.
25. Van Houwelingen HC, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis (letter). *Statistics in Medicine* 1999; **18**:107-113.
26. Gilks W, Thomas A, Spiegelhalter D. A language and program for complex Bayesian modelling. *Statistician* 1994; **43**:169-177.

5

Multivariate random-effects meta-analysis of ROC curves

Abstract

Meta-analysis of ROC-curve data is often done with fixed effects models, which suffer many shortcomings. Despite some random effects models have been published to execute a meta-analysis of ROC-curve data, these models are not often used in practice. More straightforward modelling techniques for multivariate random effects meta-analysis of ROC-curve data are needed, which can be fitted with standard software.

In this chapter sensitivities and specificities are analysed simultaneously using a bivariate random effects model. Summary ROC curves can be obtained through different characterisations of the estimated bivariate normal distribution. Under an extra assumption the model could be seen as random effects model for individual study curves. The authors fit random intercept models with approximate and with exact likelihood. Finally they extend the models to a random intercept and slope model.

The authors brought the statistical meta-analysis of ROC curve data back into a framework of relatively standard multivariate meta-analysis with random effects. The random intercept model that they propose solves the shortcomings of current fixed effects methods and is very easily fitted in practice using standard statistical software. The syntax in the software package SAS (Proc NLMIXED) that is used throughout this chapter is given in the appendix. With this syntax the bivariate random effects model is easily accessible to meta-analysists.

1 Introduction

For a thorough assessment of the effectiveness of a specific treatment, it is common to execute a meta-analysis of randomised clinical trials reported in the literature. The same is done for the assessment of the characteristics of a diagnostic test to distinguish patients having a certain disease from patients not having that disease. Meta-analyses to assess the reliability, accuracy and impact of diagnostic tests are essential to guide optimal test selection and the appropriate interpretation of test results[1]. However, the designs of test accuracy evaluations differ from the designs of studies that evaluate the effectiveness of treatments, which means that different criteria are needed when assessing study quality and potential for bias. Additionally, often each evaluation of diagnostic tests reports a pair of related summary statistics (for example sensitivity and specificity) rather than a single statistic, requiring alternative statistical methods for pooling study results[1]. Receiver Operating Characteristic (ROC) curves are used in studies of diagnostic accuracy to depict the pattern of sensitivities and specificities observed when the performance of the test is evaluated at several different diagnostic thresholds.

In the last one and a half decade a number of statistical articles on meta-analysis of diagnostic test accuracy have been written[2-11]. The methods to be used depend on the type of data that is available from the different studies. Some of the references[6, 8] discuss methods for the situation where all the individual patient data of the studies are available. Some articles discuss the situation where each study provides an estimate of the area under the ROC curve and how to combine them[12]. Other references discuss the situation where per study only one estimated pair of sensitivity and specificity (corresponding to one or more different diagnostic thresholds) is available. In this chapter we focus on this last situation, which is by far the most common in practice. The aim of the meta-analysis is to estimate the overall ROC curve of the (continuous) diagnostic marker.

If for each study only one estimated pair of sensitivity and specificity is available, the simplest approach to combine evidence about binary valued diagnostic tests is to take the average of estimates of sensitivity and specificity across studies[4,13]. However, this approach only makes sense when all tests use the same scoring rule or the same threshold value. This is very unlikely and difficult to check, because most studies do not explicitly report scoring rules or threshold values, but instead report only summary statistics[4]. Other approaches reduce the problem to calculating just one outcome measure like the area under the curve statistic (AUC)[12] or the diagnostic odds ratio (DOR)[14] and combine these using standard meta-analysis techniques.

Probably the most well known and most commonly used method in practice is the Summary ROC (SROC) method proposed by Littenberg and Moses[2] and Moses et al.[3]. They plotted the difference versus the sum of the logit(true positive rate) and logit(false positive rate) from each study. Then they fitted three types of regression lines (robust, unweighted and weighted) to these points. Finally they transformed the line to ROC space.

Despite the fact that the SROC method is predominantly used in practice, it has a number of serious shortcomings. The aim of this chapter is to present an approach that extends the SROC method, addresses its drawbacks and is still easily carried out in practice using familiar statistical packages like SAS. The method follows the general multivariate approach as described in van Houwelingen et al.[15] and Arends et al.[16,17].

In section 2 we introduce two data sets that will be used as examples. In section 3 we give an overview of the SROC method and discuss its shortcomings. In section 4 we shortly discuss other methods proposed in the literature. In section 5 the new approach is presented. In section 6 the methods are applied on the two example data sets and the results are presented. We used the SAS procedures Proc Mixed and Proc NLMixed for the analyses. The syntax that was used is given in the appendix. Finally we end with a discussion in Section 7.

2 Data examples

To illustrate the methods discussed in this chapter, we apply them to two meta-analysis data sets, one relatively small (29 studies) data set and one large data set (149 studies).

2.1 Example 1: FNAC of the Breast[18]

Giard and Hermans[18] present 29 studies evaluating the accuracy of fine-needle aspiration cytologic examination (FNAC) of the breast to assess presence or absence of breast cancer. FNAC provides a non-operative way of obtaining cells for establishment of the nature of a breast lump and therefore plays a pivotal role in the preoperative diagnostic process[18-21]. The sensitivity and specificity of FNAC were determined for each study. Sensitivity was defined as the probability of a malignant or suspect test result in patients with cancer. Specificity was defined as the probability of absence of abnormal cells in the non-patients[18]. Table 1 shows the frequencies of the FNAC outcomes given the final diagnosis of benign or malignant breast disease.

Table 1. Example 1: data from clinical studies on patients with a breast mass who underwent a fine-needle aspiration cytological examination (FNAC). Patients are cross-classified according to their final diagnosis (benign or malignant breast disease) and their FNAC result.

	FNAC results for patients with benign disease			**FNAC results for patients with malignant disease**		
Study	False Pos (Y_0)	True Neg	Total (n_0)	True Pos (Y_1)	False Neg	Total (n_1)
1	70	939	1009	979	89	1068
2	3	163	166	51	22	73
3	55	894	949	1569	152	1721
4	25	259	284	35	15	50
5	4	121	125	59	12	71
6	18	216	234	56	4	60
7	602	3117	3719	329	39	368
8	10	213	223	125	17	142
9	88	499	587	211	63	274
10	0	31	31	49	1	50
11	26	643	669	336	178	514
12	147	746	893	210	42	252
13	5	25	30	16	3	19
14	16	356	372	258	53	311
15	9	107	116	56	18	74
16	16	112	128	162	28	190
17	6	112	118	116	13	129
18	99	145	244	65	12	77
19	5	78	83	94	10	104
20	0	70	70	26	4	30
21	28	136	164	1318	249	1567
22	55	539	594	569	120	689
23	1	287	288	46	16	62
24	13	76	89	64	6	70
25	1	104	105	39	4	43
26	16	426	442	132	20	152
27	17	161	178	470	22	492
28	25	200	225	28	4	32
29	43	22	65	42	3	45

The true positive rate TPR, or sensitivity, is estimated for a study by Y_1/n_1, and the false positive rate FPR, which is 1 minus the specificity, by Y_0/n_0. See Figure 1a for a plot of the estimated TPRs against the estimated FPRs and Figure 1c for the estimated TPRs and FPRs on the logit scale.

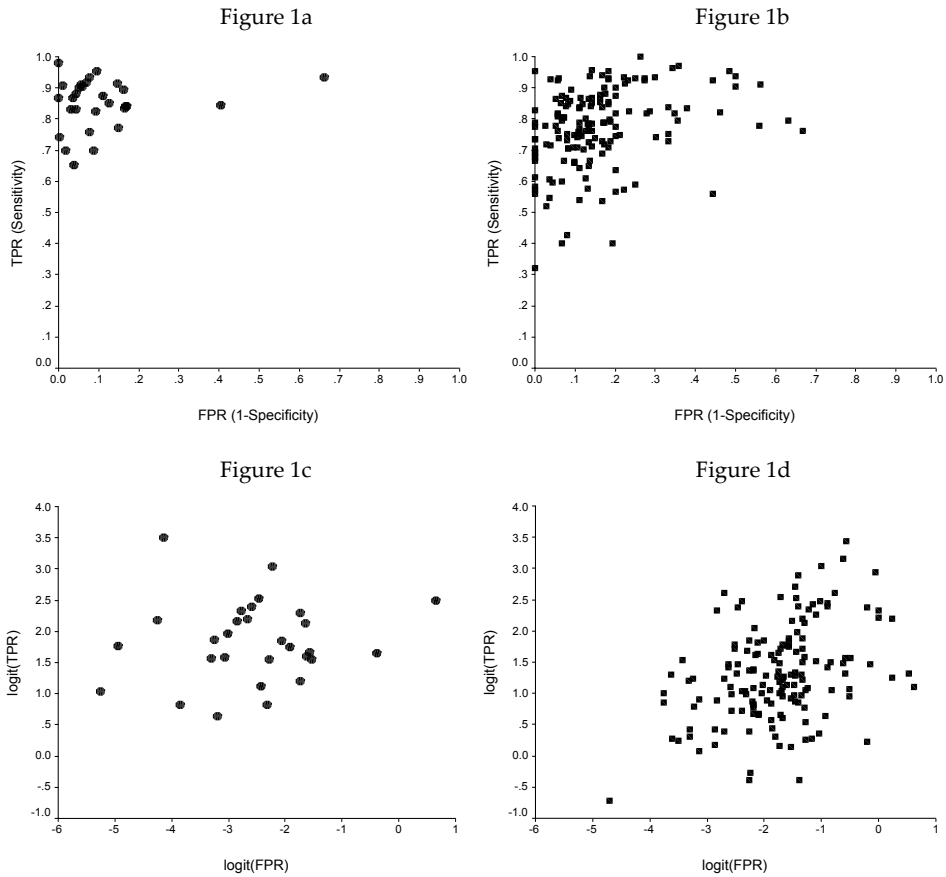


Figure 1. Observed sensitivity against (1-specificity) of data reported across 29 studies that were originally meta-analysed by Giard and Hermans[18] (left side of picture) and across 149 studies that were originally meta-analysed by Heijenbrok-Kal et al.[22] (right side of picture) on the original scale and on logit transformed scale.

The estimated TPRs and FPRs vary considerably across studies. Also, the proportions of patients with benign or malignant disease according to the final diagnosis differed substantially. At the time of publication (1992), no reasonable methods to summarize diagnostic test data across several studies were available. In this chapter we will use the data to fit the standard fixed effects SROC model as well as the proposed random effects models.

2.2 Example 2: Imaging tests for coronary artery disease[22]

Heijenbrok-Kal[22] searched PubMed from January 1990 through May 2003 for meta-analytic studies on the diagnostic performance of imaging tests for coronary artery

disease. In all meta-analyses included in this paper, angiography is the reference standard and the source numbers of true and false positives and true and false negatives are reported. Duplicate source studies are excluded. Heijenbrok-Kal[22] combined data from seven meta-analyses with a total of 246 patient series including 24,761 patients who underwent eight different imaging technologies for coronary artery disease. Coronary tests showed little difference in diagnostic performance.
For the illustration of our approach we choose from all seven selected meta-analyses only the 149 source studies in which the performance of an exercise or stress echo was investigated. These 149 studies include 13,303 patients. In Figure 1b a plot is given of the estimated TPRs against the estimated FPRs and Figure 1d represents the estimated TPRs and FPRs on the logit scale.

3 The standard SROC method

The starting point of a meta-analysis of ROC curve data is a number of studies providing information on a continuous diagnostic marker or variable *M*. In the different studies possibly different thresholds for *M* are used to obtain a dichotomous diagnostic test. The data provided by each study are the number of patients with a positive test result (y_1), the total number of patients (n_1) in the group with the disease, the number of patients with a positive test result (y_0) and the total number of patients (n_0) in the group without the disease. The aim is to estimate the overall ROC curve of the diagnostic marker M based on the available data from the different studies. The standard method used in practice is the SROC method of Littenberg and Moses[2], which proceeds as follows. The underlying model assumes that there exists a transformation of the continuous diagnostic variable *M* such that the transformed test, X, follows a logistic distribution both in the population without the disease and in the population with the disease. In other words, it is assumed that the transformation that makes the distribution of *M* logistic in the non-diseased (which always exists) makes the distribution simultaneously logistic for those with the disease. We assume that large values of X correspond with the diseased population. When small values of X correspond with the diseased population we take –X. The cumulative distribution of X in the healthy and the diseased is given by

$$\Pr(X < x \mid \text{healthy}) = \frac{e^x}{1+e^x} \quad \text{and} \quad \Pr(X < x \mid \text{disease}) = \frac{e^{-\alpha+\beta x}}{1+e^{-\alpha+\beta x}} \qquad (1)$$

for some values of $\alpha \geq 0$ and $\beta > 0$. The difference between the mean value in the population with the disease and without the disease is α/β, and the ratio between the standard deviation of the diseased and the healthy population is $1/\beta$. Thus $0<\beta<1$

corresponds with a higher variance in the population with the disease and β>1 with a smaller variance. Figure 2 gives a graphical illustration with the interpretation of α and β, where clearly 0<β<1.

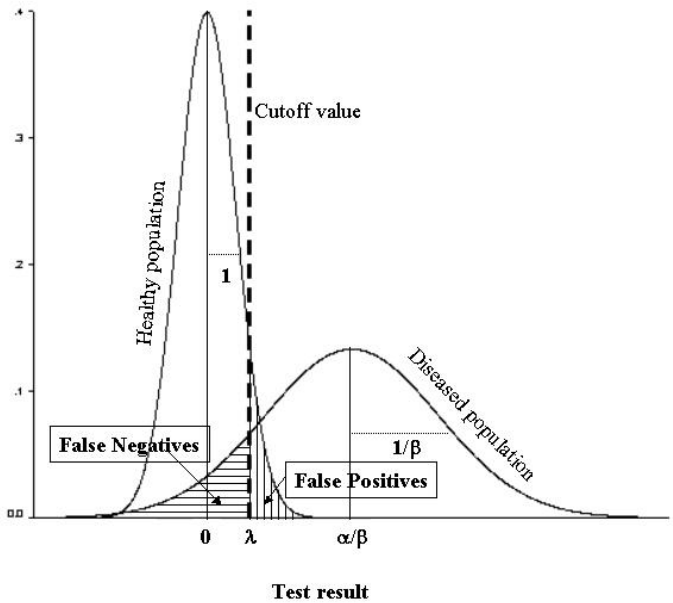


Figure 2. Graphical illustration with interpretation of α and β

If λ denotes the threshold X-value for the test being declared positive, then according to (1) the probability of a false positive result is $1- e^{\lambda}/(1+e^{\lambda})$ and hence the logit(FPR)=$-\lambda$. Similarly we have logit(TPR)=$\alpha-\beta\lambda$. In the following we will use the notation:

$$\xi = \text{logit}(FPR) = -\lambda$$
$$\eta = \text{logit}(TPR) = \alpha - \beta\lambda$$

This implies the linear relationship

$$\eta = \alpha + \beta\xi \qquad (2)$$

Following Rutter and Gatsonis[9], α can be called the accuracy parameter and β the scale or asymmetry parameter. If β =1, the resulting ROC curve is symmetric (with respect to the minus 45° diagonal), otherwise it is asymmetric.
In the SROC approach of Littenberg & Moses the relation (2) is written as

$$\eta - \xi = \alpha' + \beta'(\eta + \xi)$$

with $\alpha'= 2\alpha/(\beta+1)$ (with $\alpha \geq 0$) and $\beta'=(\beta-1)/(\beta+1)$ (with $-1<\beta'<1$). If D and S are the estimated values of $\eta-\xi$ and $\eta+\xi$ from a study (to avoid division by zero, 0.5 is added to all numbers in the 2x2 table of a study), then approximately

$$D = \alpha' + \beta' S \quad (3)$$

and the values of α' and β' are estimated by a simple weighted or unweighted linear regression. The weights are chosen proportional to the inverse variance of D. D is interpreted as the log odds ratio of a positive test result for diseased individuals relative to healthy individuals, and is often called the diagnostic odds ratio. Its estimated variance is

$$\frac{1}{y_0+0.5} + \frac{1}{n_0-y_0+0.5} + \frac{1}{y_1+0.5} + \frac{1}{n_1-y_1+0.5} \quad (4)$$

The summary ROC curve is obtained by back transforming the estimate of (3) to the ROC space. A value of $\beta' \neq 0$ indicates that the curve is asymmetric.

The advantage of the SROC method, which explains its popularity, is that it is very simple to understand and can be carried out in any statistical package. Despite this important advantage of simplicity, a number of critical comments can be made.

First of all, the SROC method is a fixed effects method, i.e. it assumes that the values of α and β do not vary across studies. Thus variation is due only to the threshold effect and within-study sampling variability. However, in many practical cases it is clear or likely that there is between study variation. Study characteristics such as technical aspects of the diagnostic test, patient selection, study settings, experience of readers etc. are among the potential contributors to between-studies variation in the estimates of diagnostic performance[9]. Modern meta-analytical methods take possible variation across studies into account by introducing random effects[15,23-26]. If there is between-studies variation, a fixed effects model can give biased estimates and typically underestimates standard errors.

Second, the independent variable S in the regression equation (3) is measured with measurement error which should be taken into account. As a result, regression to the mean[27] and attenuation due to measurement errors[28] could seriously bias the slope of the regression line[15]. Thus not taking into account the measurement error in S lead to bias in β' (in general towards zero) and α' and therefore also in β (in general towards one) and α[26].

Third, D and S are correlated within a study, positively or negatively depending on the study. In the standard fixed effects SROC model this correlation is ignored. Although probably the correlation mostly is small in practice, it is not obvious what the consequence of ignoring it is.

Fourth, it is reasonable that the different studies should be somehow weighed in the analysis, in particular if the studies vary substantially in size. If there is more than

only within study sampling variation, weighting by the inverse within study variances as is done in the weighted SROC approach will not be optimal.
Finally, to avoid undefined log odds, log odds ratios and their variances, quite arbitrarily 0.5 is added to the numbers in the fourfold tables of the trials. As Moses et al.[3] showed, the effect of this adjustment can be surprisingly large. Adding 0.5 to all cells tends to push an estimated ROC curve away from the desirable northwest corner of ROC space. The standard SROC method has to do this because it does not use the true binomial distribution of the number of positive test results within a group. It would be preferable to get rid of this artificial and arbitrary correction.
In section 5 we present a method that does not have the disadvantages of the SROC method and can still be carried out easily in standard statistical packages. But first we discuss in section 4 some other methods proposed in the literature.

4 Other methods proposed in the literature

Kardaun and Kardaun[29] also assume model (1) and exploit the linear relationship $\eta_i = \alpha+\beta\xi_i$ where i = 1, …, k denotes the number of the study. Using straightforward approximate likelihood methods all k+2 parameters (including the ξ_i's) are estimated. The estimation method is called *approximate* likelihood, since, instead of the exact likelihood based on the true distribution of the estimated ξ ($\hat{\xi}$) and η (or $\hat{\eta}$), an approximate likelihood based on the familiar normal approximations $\hat{\xi} \cong \mathrm{N}(\xi, 1/y_0 + 1/(n_0 - y_0))$ and $\hat{\eta} \cong \mathrm{N}(\eta, 1/y_1 + 1/(n_1 - y_1))$ is used. The drawbacks of the method of Kardaun and Kardaun[29] are, first, that the number of estimated parameters is proportional to the number of trials, hence standard likelihood theory does not apply. For instance, consistency of the estimates when the number of studies tends to infinity is not guaranteed. Second, their computer-intensive method based on profile likelihood is not very practical. Also the first and last drawbacks mentioned in the previous section for the SROC method still apply.
Recently Rutter and Gatsonis[9] proposed a hierarchical Bayesian regression approach, that does not have the drawbacks mentioned in the previous section for the SROC method. They assumed the following model. Let π_{i0} be the true FPR in the non-diseased and π_{i1} be the true TPR in the diseased populations. Then Y_0 ~ Binomial(n_0, π_{i0}) and Y_1 ~ Binomial(n_1, π_{i1}). Defining ξ_i = logit(π_{i0}) and η_i = logit(π_{i1}), the following relationship is assumed to hold between ξ_i and η_i:

$$\begin{aligned} \xi_i &= (\theta_i + \alpha_i X_0) e^{-\beta X_0} \\ \eta_i &= (\theta_i + \alpha_i X_1) e^{-\beta X_1} \end{aligned} \qquad (5)$$

where X_0 and X_1 are chosen to be –1/2 and +1/2 respectively. This implies the linear relationship

$$\eta_i = \alpha_i e^{-\beta/2} + e^{-\beta}\xi_i \qquad (6)$$

For equation (6), α is called the accuracy parameter, because it measures the difference between TPR and FPR, and β is called the scale parameter. With this parameterisation, if $\beta \neq 0$ the ROC curve is asymmetric.

The between study variation is modelled by assuming that α_i and θ_i are independent and normally distributed:

$$\begin{pmatrix} \alpha_i \\ \theta_i \end{pmatrix} = \mathrm{N}\left(\begin{pmatrix} \overline{\alpha} \\ \overline{\theta} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\theta^2 \end{pmatrix} \right) \qquad (7)$$

To compute a summary ROC curve, Rutter and Gatsonis[9] plug in the estimates for $\overline{\alpha}$ and β into the linear relation (6) and transform it into the ROC space.

The method allows for between-study variation by modelling the accuracy parameter α with a random effect. A minor drawback of the method is it does not allow between studies variation in the scale parameter β. Rutter and Gatsonis[9] remark that allowing the scale parameter β to vary across studies would make some model parameters unidentifiable. A more serious and practical disadvantage is that Rutter and Gatsonis[9] compute the estimates in a Bayesian way using Markov Chain Monte Carlo (MCMC) simulation with the BUGS software, which is rather complicated. MCMC estimation requires programming, simulation, evaluation of convergence and model adequacy, and synthesis of simulation results. Implementation of MCMC simulation entails non-trivial analysis tasks including evaluation of convergence and the adequacy of prior distributions, and these tasks require statistical expertise. As the authors mention, this is a high price that has to be paid for the advantages of the hierarchical SROC model. Furthermore, Rutter and Gatsonis[9] use a relatively complicated parameterisation, which can make it difficult for the meta-analyst to fully understand what he is doing. Macaskill[30] shows how the model of Rutter and Gatsonis can be fitted in a non-Bayesian way using the SAS NLMixed program for generalised linear mixed models. This makes the model of Rutter and Gatsonis model much more practical.

Recently a straightforward random effects extension of the method of Littenberg and Moses[2] has been used in some medical applications[31-33], with results obtained from the STATA program Metareg[34]. This method is as follows. Let ξ_i and η_i again denote the true logit(TPR) and logit(FPR) for study i. Let $D_i = \eta_i - \xi_i$ be the true log odds ratio and $S_i = \eta_i + \xi_i$. The corresponding estimates are given by $\hat{\xi}_i$, $\hat{\eta}_i$, $\hat{D}_i$ and $\hat{S}_i$, respectively.

Then the model is:

$$\hat{D}_i = \alpha_i + \beta \hat{S}_i$$

$$\text{with } \hat{D}_i \simeq \mathrm{N}(D_i, \tfrac{1}{y_0+0.5} + \tfrac{1}{n_0-y_0+0.5} + \tfrac{1}{y_1+0.5} + \tfrac{1}{n_1-y_1+0.5}) \text{ and } \alpha_i \simeq \mathrm{N}(\overline{\alpha}, \sigma_\alpha^2) \qquad (8)$$

In this model, all studies have a common slope β, but the intercepts vary randomly between studies according to a normal distribution. The overall ROC line is $\eta = \overline{\alpha} + \beta\xi$, where the individual study lines vary randomly around this line with between studies standard deviation σ_α. This is the standard random effects meta-regression model and there are many programs available for fitting this model. Measurement error of $\hat{D}_i$ is correctly accounted for, the measurement error in $\hat{S}_i$ is still neglected. Another drawback for sparse data sets is that it is not simply possible to use the underlying binomial distributions for $\hat{D}_i$ and $\hat{S}_i$ instead of the normal approximations.

The aim of this chapter is to present a practical method that addresses the drawbacks of the SROC method mentioned in the previous section, allows between study variation both in accuracy *and* scale parameter, and is easy to carry out in practice. The method follows the general multivariate approach as described in van Houwelingen et al.[15,35] and Arends et al.[36]. It can be implemented using standard statistical packages.

5 Alternative approach

In numerous medical articles sensitivities or specificities are meta-analysed separately by the standard random effects model of DerSimonian and Laird[23]. The method we propose is a direct extension of this approach. We analyse sensitivities and specificities simultaneously using a two-dimensional random effects model. We will show that the model implies a linear relationship between η and ξ, and can be seen as an extension of the SROC method of Littenberg and Moses[2]. We show that under an extra assumption the model can be interpreted as a random intercept model analogous to (8) that describes individual study lines as parallel lines around an overall mean line. In section 5.1 we introduce our model. In section 5.2 we discuss several types of summary ROC curves. In section 5.3 we discuss the relation with the approach of Rutter and Gatsonis[9]. In section 5.4 we take a closer look at the interpretation of our model in terms of individual study ROC curves. In the last subsection the model is generalised to allow a random slope (or scale parameter) β as well. Throughout we follow a two-level hierarchical modelling approach, explicitly modelling the within and between studies variability.

5.1 The bivariate model

The standard way of meta-analysing false positive rates of a diagnostic test in the medical literature is the DerSimonian and Laird[23] random effects model:

$$\xi_i \simeq \mathrm{N}(\overline{\xi}, \sigma_\xi^2) \text{ with } \hat{\xi}_i \cong \mathrm{N}\left(\xi_i, \frac{1}{x_0} + \frac{1}{n_0 - x_0}\right)$$

Here $\hat{\xi}_i$ and ξ_i are the observed and true logit(FPR)of study *i*, respectively. Note the well-known formula for the standard error of an estimated log odds. The parameter $\overline{\xi}$ describes the overall mean logit false positive rate and σ_ξ^2 describes the between studies variance in true logit false positive rates. Similarly, true positive rates are analysed using the model:

$$\eta_i \simeq \mathrm{N}(\overline{\eta}, \sigma_\eta^2) \text{ with } \hat{\eta}_i \cong \mathrm{N}\left(\eta_i, \frac{1}{x_1} + \frac{1}{n_1 - x_1}\right)$$

The straightforward generalisation is to assume a bivariate normal model for the pair (ξ_i, η_i):

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \cong \mathrm{N}\left(\begin{pmatrix} \overline{\xi} \\ \overline{\eta} \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma_\eta^2 \end{pmatrix}\right) \tag{9}$$

Note that this model implies the standard univariate random effects meta-analysis model for the ξ_i and η_i separately, but now allows that ξ_i and η_i are correlated.

This model fits in the framework of bivariate meta-analysis as originally introduced by van Houwelingen et al.[35]. Later on McIntosh[37] and Arends et al.[16] used this model to investigate the relationship between baseline risk and size of treatment effect in clinical trials meta-analysis. In van Houwelingen et al.[15], bivariate meta-analysis was generalized to multivariate meta-analysis and it was shown how standard General Linear Mixed Model programs can be used to fit these models. An example of a tri-variate meta-analysis is given by Arends et al.[17].

The most simple characterisation of the overall accuracy of the diagnostic test would be to take the estimated $\overline{\xi}$ and $\overline{\eta}$ and transform them to the ROC space. A more accurate description would be to characterise the bivariate normal distribution by a line and transform that line to the ROC space. Note that the bivariate normal distribution implies a linear association between ξ_i and η_i. However, as will be discussed in the next section, different lines might be employed, leading to different summary ROC curves. For example, the regression line of η_i on ξ_i could be used. Standard normal distribution theory tells that the regression line of η_i on ξ_i has intercept α and slope β given by

$$\alpha = \overline{\eta} - \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}\overline{\xi} \quad \text{and} \quad \beta = \frac{\sigma_{\xi\eta}}{\sigma_\xi^2} \tag{10}$$

The residual variance of the regression, given by $\sigma^2_{\eta|\xi} = \sigma^2_\eta - \frac{\sigma^2_{\xi\eta}}{\sigma^2_\xi}$, describes the variation in the true sensitivities between studies that have the same specificity. In section 5.2 we discuss some alternative summary ROC curves.

Similarly as in the above univariate models for meta-analysing specificities and sensitivities separately we model the within study sampling variability using the fact that the estimated logit transformed FPR, $\hat{\xi}_i$, and TPR, $\hat{\eta}_i$, are independent and approximately normally distributed:

$$\hat{\xi}_i \cong \mathrm{N}\left(\xi_i, \frac{1}{x_0} + \frac{1}{n_0 - x_0}\right) \quad \text{and} \quad \hat{\eta}_i \cong \mathrm{N}\left(\eta_i, \frac{1}{x_1} + \frac{1}{n_1 - x_1}\right) \tag{11}$$

If one or more of the denominators are close to zero, 0.5 should be added to the denominators, as in (4). The equations (9) and (11) together specify a general linear mixed model (GLMM), and the parameters can be estimated by (restricted) maximum likelihood using a GLMM program. Subsequently the intercept α and the slope β of a summary line can be calculated, using for instance (10) or one of the formulas given in the next subsection if another type of summary ROC curve is preferred. Standard errors of α and β can be calculated with the delta method. Many statistical packages provide a GLMM program. We used Proc Mixed from the SAS package. The syntax is given in the appendix. Proc Mixed does not give estimates and standard errors of user defined derived parameters, thus we had to calculate the estimates of α and β by hand, though the calculations are very simple. SAS users can avoid these hand calculations, since the model can also be fitted in Proc NLMixed. This program provides estimates and standard errors of user defined derived parameters. The syntax needed for Proc NLMixed is given in the appendix. Another possibility in Proc NLMixed is to reparameterise the model in such a way that one immediately gets the estimates and standard errors for the parameters of interest.

We call the GLMM approach the *approximate* likelihood approach, because an approximate (normal) model denoted by equation (11) is used for the within study sampling variability. The practical advantage is that the model remains a GLMM, for which much software is available. The approximate likelihood approach works well for larger data sets[15]. As a rule of thumb, the requirement 'all denominators in equation (11) larger than or equal to 5' might be adopted, though this is probably too severe.

In section 5.4 we will show that the above model, under an extra assumption, can be seen as a random effects model, describing the lines of individual studies as random varying lines parallel to an overall summary line. Thus the first drawback, that it is a fixed effects model, of the SROC method as mentioned in section 3 no longer applies. Also the model does not suffer from the second, third and fourth drawbacks. The problem of measurement error (the second drawback) is avoided by assuming a

distribution for ξ_i. In general there are two ways of dealing with measurement error, the structural and the functional approach[15]. Our approach is in the spirit of the structural approach, similar to Arends et al.[16,17] and van Houwelingen et al.[15], which has the important advantage that the parameters can be estimated by straightforward likelihood methods.

The third drawback does not apply, because $\hat{\xi}_i$ and $\hat{\eta}_i$ are independent within studies. Even if we would formulate the model in terms of D and S, as is done in the standard SROC method, then there would be no problem since the correlation can be easily modelled in the GLMM.

The fourth drawback does not apply since the likelihood method implicitly uses the 'correct' weighting based on within- as well as between-study variation. The fifth drawback still applies, since we assumed an approximate within study model. If we want to address this drawback as well, the true distribution of $\hat{\xi}_i = Y_{0i} / n_{0,i}$ and $\hat{\eta}_i = Y_{1i} / n_{1,i}$ should be used. Given the true $FPR_i = (1+\exp(-\xi_i))^{-1}$ and $TPR_i = (1+\exp(-\eta_i))^{-1}$ of study i, the observed test positive numbers Y_{0i} in the healthy group and Y_{1i} in the diseased group follow binomial distributions:

$$Y_{0i} \cong \text{Binomial}(n_{0,i}, \text{FPR}_i) \; ; \; Y_{1i} \cong \text{Binomial}(n_{1,i}, \text{TPR}_i) \qquad (12)$$

The equations (9) and (12) together now specify a General*ised* Linear Mixed Model. This model has the advantage that the fifth drawback no longer applies, but a practical disadvantage is that software for General*ised* Linear Mixed Models is not available in many packages. We again used Proc NLMixed of SAS. A syntax example is given in the appendix. We call this the *exact* likelihood approach, since the likelihood is based on the exact (i.e. binomial) within-study distribution of the data.

5.2 Choice of summary ROC curve

Above we have seen that a summary ROC curve can be obtained through a characterisation of the estimated bivariate normal distribution given by (9). One possibility is to take the regression line of η_i on ξ_i, as we did above. However, there are other possibilities as well. For example, we could take the regression line of ξ_i on η_i. We now discuss this and other possible choices.

1. *The regression line of* η_i *on* ξ_i

$$\eta = \overline{\eta} + \frac{\sigma_{\xi\eta}}{\sigma_{\xi}^2}(\xi - \overline{\xi}) \qquad (13)$$

This summary line estimates the mean logit transformed sensitivity given a specific value for the logit transformed 1-specificity. When transformed to the ROC space, the summary ROC curve estimates the median TPR given a specific value for the FPR.

2. *The regression line of ξ_i on η_i*

$$\eta = \overline{\eta} + \frac{\sigma_\eta^2}{\sigma_{\xi\eta}}(\xi - \overline{\xi}) \qquad (14)$$

This summary line characterises the mean logit transformed 1-specificity given a specific value for the logit transformed sensitivity. When transformed to the ROC space, the summary ROC curve characterises the median FPR given a specific value for the TPR.

3. *The regression line of D_i on S_i*

Let $D_i = \eta_i - \xi_i$ and $S_i = \eta_i + \xi_i$, as in the classical SROC method. From (9) it follows that the covariance of D and S is equal to $\sigma_\eta^2 - \sigma_\xi^2$ and the variance of S is equal to $\sigma_\eta^2 + \sigma_\xi^2 + 2\sigma_{\xi\eta}$. The regression line therefore is

$$D = \overline{D} + \frac{(\sigma_\eta^2 - \sigma_\xi^2)}{(\sigma_\eta^2 + \sigma_\xi^2 + 2\sigma_{\xi\eta})}(S - \overline{S})$$

The popularity of this summary line is possibly explained by the fact that it has an appealing interpretation. Given S, which can be interpreted as a proxy for the positivity criterion of the diagnostic test, this regression line estimates D, which can be interpreted as the diagnostic log odds ratio.

In terms of η and ξ the regression line is

$$\eta = \overline{\eta} + \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}(\xi - \overline{\xi}) \qquad (15)$$

This method is a kind of compromise between the vertical way of looking in the first method (median TPR given a specific value for the FPR) and the horizontal way of looking in the second method (median FPR given a specific value for the TPR). When back transformed to the ROC space, the summary ROC curve characterises the median when one looks along a line with constant $\eta+\xi$ (a minus 45° degrees line).

4. *The Rutter and Gatsonis[9] summary ROC curve*

Their method leads to the summary line (see section 5.3)

$$\eta = \overline{\eta} + \frac{\sigma_\eta}{\sigma_\xi}(\xi - \overline{\xi}) \qquad (16)$$

This line can be interpreted as a sort of compromise between the regression of η_i on ξ_i and that of ξ_i on η_i, since the slope is equal to the geometric mean of the slopes of the two regression lines.

5. *The major axis method*

The last possibility we mention is to characterise the bivariate normal distribution between ξ and η by the major axis that runs through the extreme points of the ellipses which are defined by the (1-α)100% coverage intervals of the estimated bivariate distribution. This results in the summary line

$$\eta = \overline{\eta} + \frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{\left(\sigma_\eta^2 - \sigma_\xi^2\right)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}(\xi - \overline{\xi}) \qquad (17)$$

In fact, taking this line is analogous to summarising a two dimensional distribution by it first principal component.

The summary ROC curves of methods 3-5 are symmetric in ξ and η; that is, if the labelling of diseased and non-diseased is interchanged, the summary ROC curve does not change. For all of the mentioned summary lines, standard errors for the slope, intercept and for η at a given value for ξ can be calculated using the delta method. Confidence intervals for the slope and intercept, and a confidence band for the summary line, are calculated using standard methods. A confidence band for the summary ROC curve is obtained by transforming the confidence band of the summary line. No extra programming or hand calculations are needed if a program like SAS Proc NLMixed is used that allows user defined derived parameters.

5.3 Relationship with model of Rutter and Gatsonis

From (5) and (7) it follows that the model of Rutter & Gatsonis can be written as

$$\begin{pmatrix}\xi_i\\ \eta_i\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}(\overline{\theta} + X_0\overline{\alpha})e^{-X_0\beta}\\ (\overline{\theta} + X_1\overline{\alpha})e^{-X_1\beta}\end{pmatrix}, \begin{pmatrix}(\sigma_\theta^2 + X_0^2\sigma_\alpha^2)e^{-2X_0\beta} & (\sigma_\theta^2 + X_0X_1\sigma_\alpha^2)e^{-(X_0+X_1)\beta}\\ (\sigma_\theta^2 + X_0X_1\sigma_\alpha^2)e^{-(X_0+X_1)\beta} & (\sigma_\theta^2 + X_1^2\sigma_\alpha^2)e^{-2X_1\beta}\end{pmatrix}\right)$$

This specifies a bivariate normal distribution for (ξ_i, η_i), just as we do in (9). Note that the number of parameters is the same, too. Thus the two models are essentially the same, only the parameterisation is different. Rutter and Gatsonis[9] choose X_0 = -½ and X_1 = ½ and do not discuss other choices. One can check that their labelling also lead to σ_η/σ_ξ as the slope given by (16). All other choices such that X_0 = -X_1 also lead to σ_η/σ_ξ. Alternative choices for X_0 and X_1 lead to other summary lines. For instance, the choice X_0 = 0 and X_1 = 1 leads to the η on ξ regression line given by (13). The choice X_0 = 1 and X_1 = 0 leads to the ξ on η regression given by (14). One can show that it is not possible to specify X_0 and X_1 such that it leads to the D on S regression line (15).

We conclude that our bivariate model is in principle identical to that of Rutter and Gatsonis[9]. A minor difference is the different parameterisation. Another minor difference is that the slope in the Rutter and Gatsonis model is e^{β}, and this it is restricted to be positive. We do not restrict the slope in our model, although in practice negative slope estimates will typically not occur. An important practical difference is that Rutter and Gatsonis follow a laborious Bayesian estimation approach, while our method can be carried out conveniently using standard statistical packages. Furthermore, we think our method is more straightforward and easier to understand, since it simply assumes a standard random effects model for the sensitivities and specificities simultaneously. In section 5.5 we extend the model with an extra random effect for the slope.

5.4 The bivariate model as a random intercept model

The bivariate model as introduced in section 5.1 did not assume anything about study specific curves. The method simply lead to an estimated underlying bivariate distribution of the true sensitivities and specificities. This distribution could then be characterised by a summary ROC curve, and different choices for it were available. The summary ROC curve that is chosen does not necessarily correspond with the true curves of the studies. The true study specific curves might have a substantially different slope, or there might be no study specific curves at all. The latter could be the case when the diagnostic test cannot be thought of as a continuous test. However, even in this case the analysis makes sense, since the method does not explicitly model study specific ROC curves and does not assume the existence of study specific curves. In this subsection we show that under an extra assumption the bivariate model can also be interpreted as a model that describes the distribution of the individual study ROC curves.

Suppose that in the (η, ξ) space the study specific ROC curves are straight lines with a common slope β. The lines of the different studies then only differ in level, characterised by the intercept α_i for study i:

$\eta = \alpha_i + \beta\ \xi$

We assume that the α_i's are normally distributed with mean $\overline{\alpha}$ and variance σ_α^2. The observations consist of an estimate $(\hat{\xi}_i, \hat{\eta}_i)$ of one pair (ξ_i, η_i) per study. To be able to estimate the parameters, we have to assume a model that describes how these pairs arise across studies. For instance we could assume that ξ_i values are drawn from a normal distribution with mean $\overline{\xi}$ and variance σ_ξ^2, independent of α_i. This means that the individual investigators in selecting their ξ_i value are not lead by the level of their line. We could also assume another known non-zero covariance between ξ_i and α_i.

However, it is not possible to estimate this covariance, since otherwise the model would be overparametrised. Notice that the model assumptions lead to a bivariate normal distribution for (ξ_i, η_i). In fact, we have the model introduced in section 5.1 back, only with a different parameterisation. It is easy to check that $\beta = \sigma_{\xi\eta} / \sigma_\xi^2$, thus the average line leading to the summary ROC curve is the η on ξ regression line. We see that we now can interpret the model as a random intercept regression model with average line $\eta = \bar{\alpha} + \beta\xi$ and random varying parallel study specific lines $\eta = \alpha_i + \beta\xi$. The variation between the study specific ROC curves is characterised by σ_α^2. However, we should realise that this interpretation rests on the assumption that the point on the study specific line for which we observe an estimate, is chosen independently from the level α_i. It is an untestable assumption and might be questionable.

Another possible assumption would be that the y-coordinate of the point (ξ_i, η_i) for which we observe an estimate has a normal distribution with mean $\bar{\eta}$ and variance σ_η^2, independent from α_i. This again leads to the bivariate normal model for (ξ_i, η_i) with yet another parameterisation. It is easy to check that in this case $\beta = \sigma_\eta^2 / \sigma_{\xi\eta}$, thus the average line leading to the corresponding summary ROC curve is the ξ on η regression line. Again we can interpret the model as a random intercept regression model, but now with another value of the slope. Of course this interpretation rests on the untestable assumption of independence of α_i and η_i, and whether that is a reasonable assumption might be questionable.

Another assumption that could be adopted, in the spirit of the Littenberg & Moses method[2], is that the sum of the x- and y-coordinate of the point (ξ_i, η_i), $S_i = \xi_i + \eta_i$, has a normal distribution with mean $\bar{s}$ and variance σ_s^2, independent from α_i. Again this induces a bivariate normal model for (ξ_i, η_i) with yet another parameterisation. One can check that in this case $\beta = (\sigma_\eta^2 + \sigma_{\xi\eta}) / (\sigma_\xi^2 + \sigma_{\xi\eta})$, the slope of the regression line of D on S. Now the model can also be interpreted as a random intercept model $D_i = \alpha'_i + \beta S_i$ in the (S,D) space. This interpretation rests on the assumption that the individual investigators selected their value S_i independent of the level of their individual line, an assumption that might be questionable.

In general, we could make the assumption that for some reasonably chosen numbers a and b, the value of $a\xi_i + b\eta_i$ is normally distributed independent of α_i. This always leads to a bivariate normal model for (ξ_i, η_i) and a β which can be expressed in the (co)variance parameters σ_ξ^2, σ_η^2 en $\sigma_{\xi\eta}$. Suppose we take a=1 and b=0, then we get the first one of the above mentioned cases, where the summary ROC line is identical to the η on ξ regression line. If we take a=0 and b=1, then we are in the case where the summary ROC line is the ξ on η regression line. If we take a=b=1, then this leads to the S on D summary ROC line.

To see how the approach of Rutter and Gatsonis[9] fits in, we rewrite their model denoted by equation (5) as:

$$2\sqrt{\beta}\theta_i = \beta\xi_i + \eta_i$$
$$\alpha_i = -\beta\xi_i + \eta_i$$

where we have replaced exp(-β) by β, and $\alpha_i$exp(-β/2) by α_i. The study specific ROC lines are then expressed as $\eta = \alpha_i + \beta\ \xi$. Rutter and Gatsonis[9] assume that θ_i is independent from α_i. It follows they assume that aξ_i+bη_i with a = β and b = 1 is independent from α_i. One can easily check that in this case $\beta = \sigma_\eta/\sigma_\xi$, the value already mentioned in subsection 5.3. In the (α, θ) space, the study specific lines have slope zero, therefore the independence assumption may not be unreasonable. However, the assumption remains untestable. Rutter and Gatsonis argue: 'The assumed conditional independence of θ_i and α_i reflects assumptions implicit in ROC analysis. In the context of ROC analysis, positivity threshold and accuracy are independent test characteristics that together impose correlation between a test's sensitivity and specificity.' We are not able to fully understand this argument, particularly not because in their article this argument precedes the choice of the labels for X_0 and X_1 (see our equation (5)). In section 5.3 we have seen that different choices of the labelling can lead to different β's. This seems to contradict their argument, because the argument cannot be true independent of the labelling.

We conclude from this subsection that the bivariate model can be interpreted as a random intercept regression model, describing the distribution of the individual ROC curves and the variation between them. However, this interpretation always rests on an intestable assumption about independence of the level of the study specific line from a certain specified linear combination of the x- and y-coordinate of the point on the ROC curve for which the estimate is provided. Once one believes the assumption being reasonable, the interpretation is allowed.

5.5 Random intercept and slope model

Each of the random intercept models mentioned in section 5.4 can be extended with an extra random effect that allows the slopes of the individual study lines to vary as well. Analogously to what we did in section 5.4, again we have to assume that for some reasonably chosen numbers a and b, the value of aξ_i+bη_i is normally distributed, independent of the study specific line, i.e. independent of α_i and β_i.

Suppose we take the a=1 and b=0, the choice that is associated with the η on ξ regression line. The model is:

$$\eta_i = \alpha_i + \beta_i \xi_i \quad with \quad \begin{pmatrix} \xi_i \\ \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{\xi} \\ \overline{\alpha} \\ \overline{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ 0 & \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix} \right) \tag{18}$$

Analogous to the bivariate model above, this model could be termed a tri-variate model, since there are three random effects involved. Note that, even with the approximate measurement error model (12), the model is no longer a GLMM, but a General*ised* Linear Mixed Model. Therefore it does not have a practical advantage anymore to use the approximate measurement error model. The exact within study measurement error model now assumes that, given (ξ_i, α_i, β_i), the true and false positive numbers Y_{ij} (with j=0/1 for non-diseased/diseased) are independent Binomial(π_{ij}, n_{ij}) distributed with

$$\begin{aligned} \log it(\pi_{io}) &= \xi_i \\ \log it(\pi_{i1}) &= \alpha_i + \beta_i \xi_i \end{aligned} \tag{19}$$

The summary ROC curve is obtained by transforming the estimate of $\eta = \overline{\alpha} + \overline{\beta}\xi$ back into the ROC space.

If one prefers the specification a=0 and b=1, corresponding to the ξ on η regression line as summary ROC curve, extension of the random intercept normal model is done analogously, by interchanging the role of η_i and ξ_i.

If one is willing to assume that ξ_i+η_i is independent from the study specific regression line, which corresponds to the D on S regression summary ROC curve, the extended model can be formulated as

$$D_i = \alpha_i + \beta_i S_i \quad \text{with} \quad \begin{pmatrix} S_i \\ \alpha_i \\ \beta_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \overline{S} \\ \overline{\alpha} \\ \overline{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_S^2 & 0 & 0 \\ 0 & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ 0 & \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix} \right) \tag{20}$$

The within study measurement error model assumes that, given (S_i, α_i, β_i), Y_{ij} (j=0,1) are independent Binomial(π_{ij}, n_{ij}) distributed with

$$\begin{aligned} \log it(\pi_{io}) &= (\alpha_i + (1+\beta_i)S_i)/2 \\ \log it(\pi_{i1}) &= (-\alpha_i + (1-\beta_i)S_i)/2 \end{aligned} \tag{21}$$

Also the Rutter&Gatsonis model can be extended with a random slope as follows.

$$\eta = \alpha_i e^{-\beta/2} + e^{-\beta}\xi \quad \text{with} \quad \begin{pmatrix} \theta_i \\ \alpha_i \\ \beta_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \overline{\theta} \\ \overline{\alpha} \\ \overline{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & 0 & 0 \\ 0 & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ 0 & \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix} \right) \tag{22}$$

The within study measurement error model assumes that, given (θ_i, α_i, β_i), Y_{ij} (j=0,1) are independent Binomial(π_{ij}, n_{ij}) distributed with

$$\begin{aligned} \log it(\pi_{io}) &= (\theta_i - \alpha_i/2)e^{\beta_i/2} \\ \log it(\pi_{i1}) &= (\theta_i + \alpha_i/2)e^{-\beta_i/2} \end{aligned} \tag{23}$$

All models in this section can be fitted with SAS Proc NLMixed. As an example we give the syntax of fitting the summary ROC curve based on the D on S regression in the appendix.
It might be surprising that it is possible to fit the model with random slopes, while there is only one point per study available. However, one should realise that under this model (ξ_i, η_i) is not longer bivariately normal, but has another non-normal bivariate distribution. It is the deviation from the normal distribution that makes the parameters identifiable. See van Montfort et al.[38] for a comparable problem.
For the random intercept model discussed in the previous subsection, the model for (ξ_i, η_i) and therefore for the observations ($\hat{\xi}_i$, $\hat{\eta}_i$) was the same, irrespective of the values of a and b. In the present random intercept and random slope models, the models become really different for different values of a and b. In practice, one might try different choices, and select the model that fits the data best.
For the bivariate model or random intercept model, except for the models with the η on ξ and ξ on η regression lines, were invariant with respect to changing of the disease and non-disease label. Unfortunately, this nice property is lost in the random intercept/random slope model.

6 Results

6.1 Random intercept model

Example 1: FNAC of the Breast[18]

We fitted the random intercept model as described in section 5.1 on the data of the 29 studies of the meta-analysis of Giard et al.[18]. The estimates of the means and variances of η_i and ξ_i resulting from the approximate and exact likelihood approach are presented in the upper part of Table 2. Based on these estimates, the results for the five different choices of the summary ROC curve (section 5.2) are presented in the lower part of Table 2.
In Figure 3 the different ROC curves are depicted, in the logit-logit space as well as in the ROC space. Also the 95% coverage regions are given. These regions are based on the fitted bivariate distribution and estimate the area that contains approximately 95% of the true pairs of (logit(FPR), logit(TPR)) and (FPR, TPR) respectively.
From Table 2 and the Figure 3 it is clear that the results of the exact and approximate approach are similar in this data example. The exact approach results in a somewhat more favourable average sensitivity and specificity.

Table 2. First data example: FNAC of the breast[18]. In the upper part estimates are given of the random intercept model (section 5.1), using approximate as well as exact likelihood. In the lower part the parameter estimates are given for the five different choices of the summary ROC curves discussed in section 5.2.

Parameter	Approximate likelihood	Exact likelihood
	Estimate (se)	Estimate (se)
mean logit(TPR) ($\overline{\eta}$)	1.774 (0.114)	1.839 (0.119)
mean logit(FPR) ($\overline{\xi}$)	-2.384 (0.201)	-2.547 (0.225)
var(logit(TPR)) (σ_η^2)	0.286 (0.093)	0.316 (0.104)
var(logit(FPR)) (σ_ξ^2)	0.990 (0.313)	1.297 (0.411)
cov(logit(TPR),logit(FPR)) ($\sigma_{\eta\xi}$)	0.146 (0.132)	0.141 (0.155)

Type of summary ROC	Approximate likelihood		Exact likelihood	
	α (se)	β (se)	α (se)	β (se)
1. η on ξ	2.126 (0.32)	0.148 (0.13)	2.115 (0.32)	0.108 (0.12)
2. ξon η	6.431 (3.95)	1.954 (1.65)	7.560 (6.13)	2.246 (2.39)
3. D on S	2.680 (0.37)	0.380 (0.15)	2.647 (0.37)	0.318 (0.14)
4. Rutter & Gatsonis	3.054 (0.31)	0.537 (0.12)	3.096 (0.32)	0.494 (0.11)
5. Major axis	2.249 (0.42)	0.199 (0.17)	2.196 (0.41)	0.141 (0.15)

This was to be expected beforehand for two reasons. First, as mentioned in section 3, adding 0.5 to the numbers in the fourfold table, as is done in the approximate approach, results in estimated mean sensitivity and specificity that are biased downwards, pushing the ROC curve away from the left upper corner. Second, as shown by Chang et al.[39], even if it is not needed to add 0.5, the estimates of the mean sensitivity and specificity are still somewhat biased towards 0.5. This is due to the fact that the approximate approach does not account for the correlation between the logit(TPR) and its variance, and between the logit(FPR) and its variance.

From Table 2 and Figure 3 it is clear that the difference among the different types of the summary ROC curve is substantial, especially for the first two choices 'η on ξ' and 'ξon η'. As one can see on the basis of the formulas for the slopes given in section 5.2, the first two types ('η on ξ' and 'ξon η') give a kind of lower and upper bound for the estimated summary ROC curves, and types 3 to 5 lie between these two curves. In fact, the slopes of choices 3 to 5 could be considered different kinds of 'weighted averages' of the slopes of methods 1 and 2.

Figure 3a

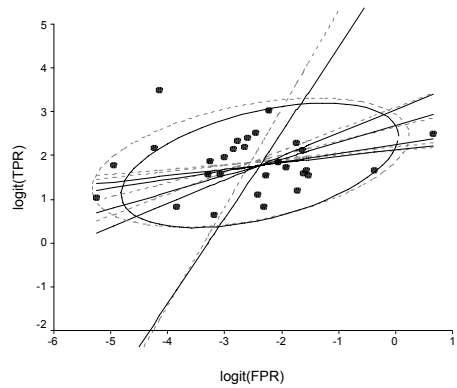


Figure 3b

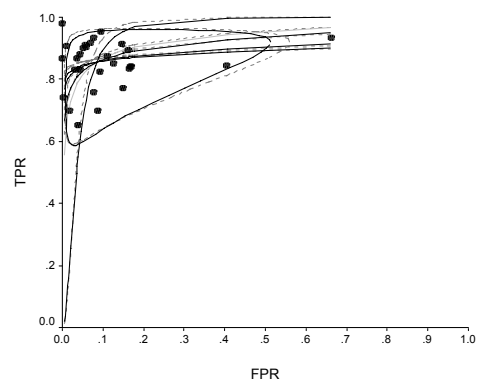


Figure 3c

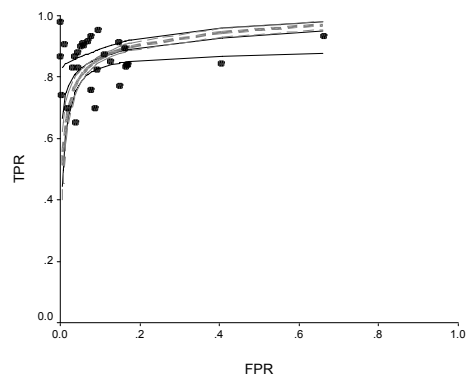


Figure 3. SROC curves for the five different choices of the summary ROC curve, as a graphical illustration of Table 2. The curves are presented in logit-logit space (Figure 3a) as well as in the ROC space (Figure 3b). Also the 95% coverage regions are given as an ellipse in Figure 3a and a 'triangle' in Figure 3b. The solid lines present the results of the approximate likelihood, the grey dashed lines present the results of the exact likelihood. In Figure 3c the SROC curves of the random intercept and slope model (solid lines) versus the fixed Littenberg & Moses model (grey dashed lines) are given together with their confidence intervals.

In this example the curves for approaches 3, 4 and 5 lie closer to the regression of η on ξ, but in general that is not the case. Results depend on the variances of ξ_i and η_i, and the covariance between them. The more similar the variances of ξ_i and η_i are, the more similar will be the results of approaches 3-5.

For all summary ROC curves given in Figure 3a and 3b a confidence band can be calculated. As an example, we have drawn in Figure 3c the 'D on S' summary ROC curve together with its confidence band. From all 5 types of summary ROC curves,

this 'D on S' summary ROC curve should be most comparable to the standard summary ROC curve from the Littenberg&Moses[2] (L&M) approach, which also estimates the regression of D on S. To compare the two, we have also drawn the L&M summary curve and its confidence band in Figure 3c. The L&M summary ROC curve has a slope that is considerably steeper than our 'D on S' curve, leading to larger estimated sensitivities if the specificity is small, and smaller estimated sensitivities if the specificity is large. This is not a general pattern, as will be seen from the second data example. Furthermore, it is seen that the L&M approach grossly underestimates the variability in the data, leading to a much too narrow confidence band. This is due to the fact that the L&M approach is based on a fixed effects model, which erroneously assumes that there is no between studies variability.

Second example: Imaging tests for coronary artery disease[22]

We fitted the random intercept model (section 5.1) on the data of the 149 studies included in the meta-analysis of Heijenbrok-Kal et al.[22]. The estimates of the means and variances of η_i and ξ_i based on the approximate and exact likelihood approach are presented in the first part of Table 3. Based on these estimates, the results for the five different choices of the summery ROC curve (section 5.2) are presented in the lower part of Table 3.

In Figure 4a and b the different summary ROC curves are given for the exact and the approximate approach, in the logit-logit space as well as in the original ROC space.

Also in this example the results of the approximate and exact likelihood approach are similar. In the approximate likelihood approach the variances of ξ and η are almost equal, which results in very little differences among the methods 3 to 5. For the exact likelihood approach the difference between the two variances is somewhat larger, leading to somewhat larger differences between the types 3 to 5. Notice that in Figure 4 considerably more than 5 percent of the studies fall outside the 95% coverage region. However, this is expected since the coverage ellipse describes the variation between the true pairs of sensitivity and specificity, while the points in the plot represent the estimates (observed) pairs of sensitivity and specificity. The observed points, of course, should show more variation due to within study sampling variability. In Figure 4c we compare again our 'D on S' summary ROC with the standard L&M one. In contrast to the previous example, now the slope of the L&M ROC is smaller than that of our 'D on S' curve. Again it is clear that L&M method leads to smaller standard errors.

As a by-product of fitting random effects model most programs provide the empirical Bayes[40,41] estimates of the study specific random effects. These can be easily used to calculate the estimates of the study specific ROC curves.

Table 3. Second data example: Imaging tests for coronary artery disease[22]. In the upper part estimates are given of the random intercept model (section 5.1), using approximate as well as exact likelihood. In the lower part the parameter estimates are given for the five different choices of the summary ROC curves discussed in section 5.2.

Parameter	Approximate likelihood	Exact likelihood
	Estimate (se)	Estimate (se)
mean logit(TPR) ($\overline{\eta}$)	1.257 (0.057)	1.339 (0.061)
mean logit(FPR) ($\overline{\xi}$)	-1.560 (0.071)	-1.851 (0.085)
var(logit(TPR)) (σ_η^2)	0.333 (0.057)	0.406 (0.066)
var(logit(FPR)) (σ_ξ^2)	0.337 (0.074)	0.585 (0.117)
cov(logit(TPR,logit(FPR)) ($\sigma_{\eta\xi}$)	0.182 (0.049)	0.272 (0.065)

Type of summary ROC	Approximate likelihood		Exact likelihood	
	α (se)	β (se)	α (se)	β (se)
1. η on ξ	2.098 (0.21)	0.540 (0.13)	2.199 (0.18)	0.465 (0.10)
2. ξ on η	4.106 (0.69)	1.827 (0.45)	4.102 (0.58)	1.493 (0.31)
3. D on S	2.802 (0.26)	0.991 (0.17)	2.802 (0.23)	0.791 (0.12)
4. Rutter & Gatsonis	2.805 (0.21)	0.993 (0.13)	2.880 (0.19)	0.833 (0.10)
5. Major axis	2.796 (0.37)	0.987 (0.24)	2.677 (0.28)	0.723 (0.15)

Empirical Bayes estimates are in a certain sense the optimal estimates of the study specific curves that take into account the estimated distribution of curves, see for instance Carlin & Louis[40]. As an example we give in Figure 5a the estimated curves of 17 studies of our example.

6.2 Random intercept and slope model

As discussed in section 5.4 the random intercept model can be extended with an extra random effect for the slope. We could fit the extended model for all 5 different choices of the summary ROC curve. As an example we will only show the results for the 'D on S' regression line in the second data example. This choice is motivated by the fact that it is more or less natural to regress the log odds ratio of a positive test result for diseased individuals relative to healthy individuals (D, sometimes called the diagnostic log odds ratio) on a kind of threshold value of the test (S).

Figure 4a

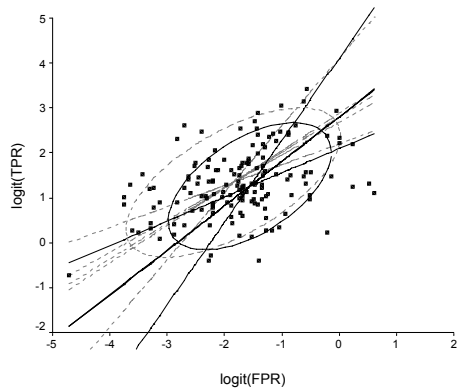


Figure 4b

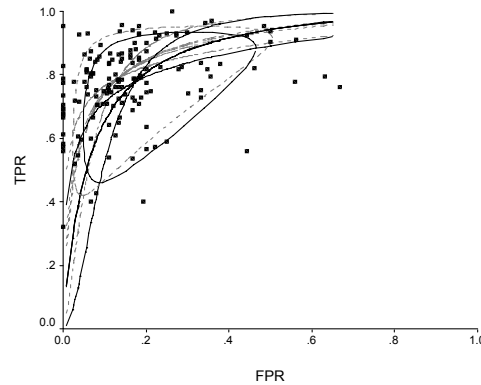


Figure 4c

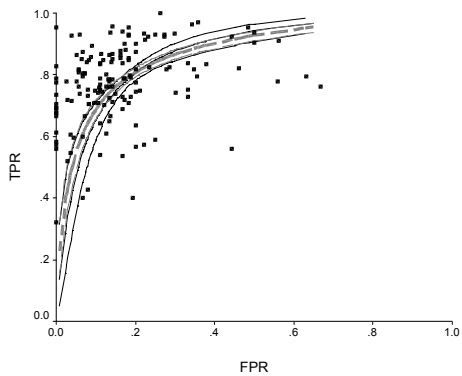


Figure 4. Second data-example of Heijenbrok-Kal et al. ROC curves for the five different choices of the summary ROC curve, as a graphical illustration of Table 3. The curves are presented in logit-logit space (Figure 4a) as well as in the ROC space (Figure 4b). Also the 95% coverage regions are given as an ellipse in Figure 3a and a 'triangle' in Figure 4b. The solid lines present the results of the approximate likelihood, the grey dashed lines present the results of the exact likelihood. In Figure 4c the SROC curves of our random intercept and slope model (solid lines) versus the fixed Littenberg & Moses model (grey dashed lines) are given together with their confidence intervals.

Moreover the 'D on S' regression line lies in between the other choices of the summary ROC curves and is never an extreme choice.

The results for the model denoted by equation (20) are given in Table 4. Note that both $\sigma_{S\alpha}$ and $\sigma_{S\beta}$ are assumed to be zero.

Figure 5a

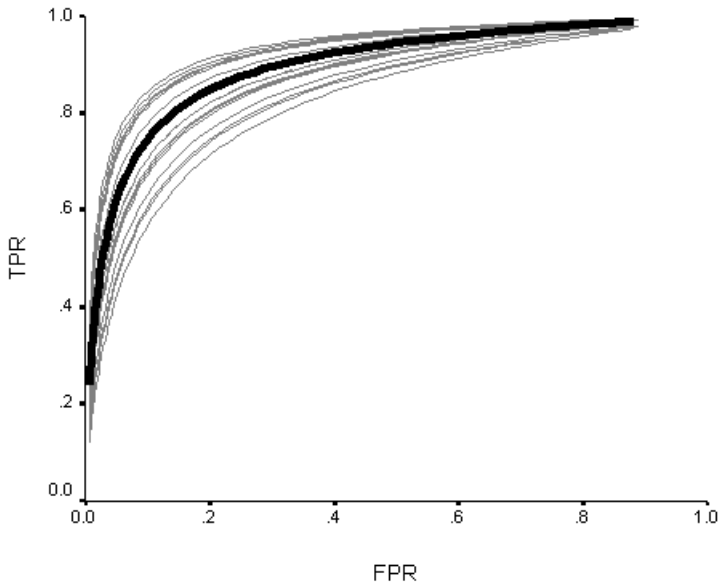


Figure 5b

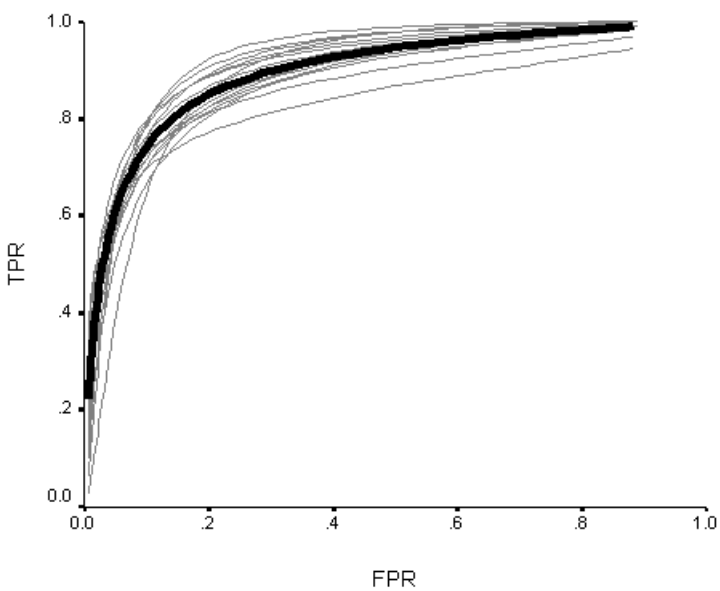


Figure 5. The Summary ROC curve (black solid line in the middle of the pictures) together with 17 study-specific ROC curves of the data-example of Heijenbrok-Kal are presented after fitting the exact random intercept model (Figure 5a) and the exact random intercept and slope model (Figure 5b).

Table 4. Results for the model denoted by equation (20) of the 'D on S' regression with random intercept and slope estimated with exact likelihood.

Parameter		Estimate (se)
mean threshold	($\overline{S}$)	-0.5012 (0.121)
mean intercept	($\overline{\alpha}$)	3.1378 (0.087)
mean slope	($\overline{\beta}$)	-0.0953 (0.087)
var(threshold)	(σ_S^2)	1.52 (0.25)
var(intercept)	(σ_α^2)	0.30 (0.13)
var(slope)	(σ_β^2)	0.10 (0.08)
cov(int, slope)	($\sigma_{\alpha\beta}$)	0.09 (0.06)

From Table 4 we can derive the following equation: D = 3.1378 - 0.0953·S, which is equivalent to η = 2.865 + 0.826·ξ. This regression line is almost equal to the 'D on S' type of summary ROC curve in Table 3, which is equal to η = 2.802 + 0.791·ξ.
We can assess with the likelihood ratio test whether the extra random effect gives a significant improvement of the model. In this example the result is: χ^2 (df=2) = 3.8, p=0.15. Hence the extra random effect gives no significant improvement of the model and could be left out, resulting in study-specific ROC-curves which will all have the same slope.
In the extended model we have two random effects per study, α_i and β_i. As a by-product SAS Proc NLMixed gives the estimates of them and we can calculate the study specific ROC curves. Since the asymmetry parameters β can differ now between studies, the study specific ROC curves can cross. This is in contrast with the random intercept model where that is not possible (see Figure 5a). This is illustrated in Figure 5.

7 Discussion

Meta-analysis of diagnostic tests requires statistical techniques that analyse pairs of related summary statistics (e.g. sensitivity and specificity) rather than a single statistic. In the literature numerous meta-analyses are published in which one is interested in meta-analysing only sensitivities or only specificities. For these situations the standard method of analysis is the DerSimonian-Laird univariate random effects model. The method we propose in this chapter is a direct extension of

that approach. We analyse sensitivities and specificities simultaneously using a two-dimensional random effects model. This model implies a linear relationship between the (logit transformed) sensitivity and specificity, which can be transformed into ROC space to obtain a summary ROC curve. We show that there are different choices for characterizing the estimated bivariate normal distribution by a regression line (i.e. the summary ROC curve). Advantages of this approach are, first, that it puts meta-analysis of ROC curve data within the framework of mainstream meta-analysis methods. Second, no assumptions about individual study curves have to be made. Third, the method does not require an underlying continuous diagnostic test, and hence it can also be applied to intrinsically dichotomous tests.

However, in section 5.4 we have shown that the bivariate model can also be interpreted as a model that describes the distribution of the individual study ROC curves and the variation between them. Under an extra independence assumption concerning how the reported points on the individual study specific ROC curves are selected, the model also provides individual study specific ROC curves. Nevertheless, this assumption remains untestable. The interpretation of the individual ROC curves rests on this assumption and interpretation is allowed once one believes the assumption is reasonable.

When using a random intercept model, we assume that the study specific ROC curves are parallel lines around the summary ROC curve. In this chapter we have shown that it is possible to relax this assumption. It might be surprising, but we show that it is also possible to fit a random slope next to a random intercept, even when there is only one point per study available. When we fit this random intercept and slope model, the study specific ROC curves of course still lie around the summary ROC curve, but are not necessary parallel to it anymore.

Our modelling approach can also be seen as an extension of the fixed effects method of Littenberg and Moses[2]. Despite a number of shortcomings, discussed in section 3, the method of Littenberg and Moses[2] still seems to be the most popular method for meta-analysis of diagnostic accuracy data where pairs of sensitivity and specificity per study are available. This is probably due to the fact that the method is very easy to carry out in practice. Although the method of Rutter and Gatsonis[9] is an appropriate alternative of the Littenberg and Moses method without its shortcomings and has been available for half a decade, we found from a literature search that this method has rarely been used. This is probably due to the fact that it is considered to be a complicated and laborious method. The method could become more popular in the future, since it was recently pointed out how this method can be performed in a non-Bayesian way using standard statistical software[30]. In this chapter we have shown how the method of Rutter & Gatsonis relates to our bivariate model.

Depending on the kind of extra independence assumption one is willing to make, different models are obtained describing individual study ROC curves and leading to different summary ROC curves. The method of Rutter & Gatsonis is just one of them.
In this chapter we are the first to put forward the issue of different types of summary ROC curves. We discussed 5 types of summary ROC curves, each of which has its own interpretation and properties. In the Littenberg and Moses approach, the choice is made explicitly as the regression of D on S. In the approach of Rutter & Gatsonis[9], the choice is implicitly made, and we pointed out that it is a kind of geometric mean between the regression line of logit(TPR) on logit(FPR) and the regression line of logit(FPR) on logit(TPR). Thus the two methods estimate different summary curves and the resulting curves are therefore in principle not the same.
We fitted our models with standard software based on straightforward likelihood methods. In our examples this approach worked well, although sometimes some convergence problems were met. In our two clinical data examples these convergence problems were encountered by specifying better starting values. However, we can imagine that, especially for small meta-analyses, this could be more of a problem. An alternative is to fit the models in a Bayesian way. This can be done using the free available software program WinBugs. The advantage is that one is very free in modelling, relaxing some model assumptions for instance. Also in applications with a relatively small number of studies, the Bayesian method might perform better, since the standard likelihood is based on large sample theory. A disadvantage is that it is more time consuming, can suffer from convergence problems, and is less easily done by non-statisticians.
In this chapter, using the framework of multivariate meta-analysis[15,35,42,43], we have shown how the analysis of ROC curve data can be performed within the framework of standard meta-analysis methods. The random intercept model that we proposed solves the shortcomings of the L&M method and is very easily fitted in practice using standard statistical software. The syntax is given in the appendix to make the model easily accessible to meta-analysists. It was surprising to discover that the model of Rutter & Gatsonis, apart from a different parameterisation, is theoretically equivalent to our bivariate meta-analysis method. The methods presented in this chapter are easily extended with covariates. If a GLMM program is used to estimate the random intercept model denoted by equation (9), both the mean logit(FPR) and mean logit(TPR) could be allowed to depend on covariates. If a Generalised Linear Mixed Model program such as SAS Proc NLMixed is used, there are many more possibilities, depending on how the model is parameterised. For instance, suppose one wishes to characterise the accuracy of the diagnostic test with

the 'D on S' regression line. Then it is possible to allow the intercept and/or the slope to depend on (possibly different sets of) covariates.
The bivariate model we proposed in this chapter can be fitted using approximate or exact likelihood. Using approximate likelihood has the advantage that a GLMM program can be used, which is widely available. The exact likelihood method can only be used if one has an appropriate General*ised* Linear Mixed Model program available. Unfortunately these programs are still rather scarce. Simulation studies are needed to compare the two approaches.

Appendix

In this appendix we provide the SAS syntax needed to reproduce the results given in Table 2 and 4. First we describe the data format that is needed. We have two records per study, one for the diseased and one for the healthy group, as in the following table.

study	group	n	npos	disease	healthy	y	est
.	.	.	.	.	.	.	0.20
.	.	.	.	.	.	.	0.10
.	.	.	.	.	.	.	0.20
1	1	1009	70	0	1	-2.58974	0.01525
1	2	1068	979	1	0	-2.58974	0.01219
2	3	166	3	0	1	-3.84405	0.29183
2	4	73	51	1	0	-3.84405	0.06386
3	5	949	25	0	1	-2.77988	0.01914
:	:	:	:	:	:	:	:

The meaning of the variables is:

`study` = number of the study
`group` = unique identifier for the diseased and healthy group
`n` = number per group
`npos` = number with positive diagnostic test
`disease` = 0 for healthy group, = 1 for diseased group
`healthy` = 1 for healthy group, = 0 for diseased group
`y` = ln(npos/(n-npos))
`est` = 1/(npos+0.5) + 1/(n-npos+0.5)

The first three lines have only a non-missing value for the variable `est`. These three values serve as starting values for the variance of ξ_i.

The following syntax produces the approximate likelihood results given in the upper part of Table 2.

```
proc mixed cl method=ml data=giardcol;
class study d group;
model y = disease healthy/noint s cl covb ddf=1000,1000;
random disease healthy / subject=study type=un s;
parms/parmsdata=giardcol eqcons=4 to 61;
repeated/group=group;
run;
```

The different summary ROC curves have to be calculated by hand based on the output of the program. The same results can also be obtained with SAS Proc NLMixed. The advantage is that the parameters of the summary ROC curves can be specified as derived parameters. The syntax is as follows.

```
proc nlmixed data=giardcol;
parms meaneta=1.8 meanksi=-2.4 vareta=0.3 varksi=1 covksieta=0.15 ;
model y~normal(eta*disease+ksi*healthy,est);
random ksi eta~normal([meanksi,meaneta],[varksi,covksieta,vareta])
subject=study;
estimate 'eta on ksi: beta' covksieta/varksi;
estimate 'eta on ksi: alpha' meaneta-covksieta/varksi*meanksi;
estimate 'ksi on eta: beta' vareta/covksieta;
estimate 'ksi on eta: alpha' meaneta-vareta/covksieta*meanksi;
estimate 'D on S: beta' (vareta+covksieta)/(varksi+covksieta);
estimate 'D on S: alpha' meaneta- (vareta+covksieta)/
(varksi+covksieta)*meanksi;
estimate 'R&G: beta' (vareta**0.5)/(varksi**0.5);
estimate 'R&G: alpha' meaneta-(vareta**0.5)/(varksi**0.5)*meanksi;
estimate 'major axis: beta' (vareta-varksi+((vareta-varksi)
**2+4*covksieta**2)**0.5)/(2*covksieta);
estimate 'major axis: alpha' meaneta-(vareta-varksi+((vareta-
varksi)**2+4*covksieta**2)**0.5)/(2*covksieta)*meanksi;
run;
```

The following syntax reproduces the right half of Table 2.

```
proc nlmixed data=giardcol;
parms meaneta=1.8 meanksi=-2.4 vareta=0.3 varksi=1 covksieta=0.15 ;
pi = 1/(1+exp(-(eta*disease+ksi*healthy)));
model npos~binomial(n,pi);
random ksi eta ~ normal([meanksi,meaneta],[varksi,covksieta,vareta])
subject=study;
estimate 'eta on ksi: beta' covksieta/varksi;
estimate 'eta on ksi: alpha' meaneta-covksieta/varksi*meanksi;
estimate 'ksi on eta: beta' vareta/covksieta;
estimate 'ksi on eta: alpha' meaneta-vareta/covksieta*meanksi;
estimate 'D on S: beta' (vareta+covksieta)/(varksi+covksieta);
estimate 'D on S: alpha' meaneta-(vareta+covksieta)/
(varksi+covksieta)*meanksi;
estimate 'R&G: beta' (vareta**0.5)/(varksi**0.5);
estimate 'R&G: alpha' meaneta-(vareta**0.5)/(varksi**0.5)*meanksi;
estimate 'major axis: beta' (vareta-varksi+((vareta-varksi)**2 +
                            4*covksieta**2)**0.5)/(2*covksieta);
estimate 'major axis: alpha' meaneta-(vareta-varksi+((vareta-varksi)**2 +
                            4*covksieta**2)**0.5)/(2*covksieta)*meanksi;
run;
```

The following syntax produces the results of Table 4.

```
proc nlmixed data=example2 df=1000;
parms malpha=2.8 mbeta=-0.1 mS=-0.5 valpha=0.4 covalbet=0 vbeta=0 vS=0.5;
  D=alpha+beta*S;
  eta=exp(((S+D)/2)*disease + ((S-D)/2)*healthy);
  pi=eta/(1+eta);
model npos ~ binomial(n,pi);
random alpha beta S ~
normal([malpha,mbeta,mS],[valpha,covab,vbeta,0,0,vS])
       subject=study;
estimate 'alpha DS' (malpha)/(1-mbeta);
estimate 'beta DS' (1+mbeta)/(1-mbeta);
run;
```

References

1. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith DG, Altman DG, editors. *Systematic Reviews in Health Care: Meta-analysis in context.* London: BMJ Publishing Group; 2001.
2. Littenberg B, Moses LE. Estimating Diagnostic-Accuracy from Multiple Conflicting Reports - a New Meta-analytic Method. *Medical Decision Making* 1993; **13**(4):313-321.
3. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993; **12**(14):1293-316.
4. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychological Bulletin* 1995; **117**(1):167-78.
5. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**(1):119-30; discussion 131-2.
6. Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of ROC curves. *Medical Decision Making* 1999; **19**(3):252-64.
7. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *Journal of Clinical Epidemiology* 1999; **52**(10):943-951.
8. Kester AD, Buntinx F. Meta-analysis of ROC curves. *Medical Decision Making* 2000; **20**(4):430-439.
9. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**(19):2865-2884.
10. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine* 2002; **21**:1237-1256.
11. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics* 2003; **59**:936-946.
12. McClish DK. Combining and comparing area estimates across studies or strata. *Medical Decision Making* 1992; **12**(4):274-279.
13. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Medical Decision Making* 1993; **13**(3):253-257.
14. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 2003; **56**:1129-1135.
15. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589-624.

16. Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24):3497-3518.
17. Arends LR, Vokó Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine* 2003; **22**:1335-1353.
18. Giard RWM, Hermans J. The Value of Aspiration Cytologic Examination of the Breast - a Statistical Review of the Medical Literature. *Cancer* 1992; **69**(8):2104-2110.
19. Mushlin AI. Diagnostic tests in breast cancer: Clinical strategies based on diagnostic probabilities. *Annals of Internal Medicine* 1985; **103**:79-85.
20. American College of Physicians. The use of diagnostic tests for screening and evaluating breast lesions. *Annals of Internal Medicine* 1985; **103**:147-151.
21. Dixon JM, Clarke PJ, Crucioli V, Dehn TCB, Lee ECG, Greenal MJ. Reduction of the surgical excision rate in benign breast disease using fine needle aspiration cytology with immediate reporting. *British Journal of Surgery* 1987; **74**:1014-1016.
22. Heijenbrok-Kal MH. *Assessment of diagnostic imaging technologies for cardiovascular disease (dissertation).* Erasmus MC: Rotterdam, 2004.
23. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177-188.
24. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* 1999; **18**(3):321-359.
25. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* 1998; **17**:841-856.
26. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 1994; **309**(6965):1351-1355.
27. Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). *Statistics in Medicine* 1994; **13**(3):293-296.
28. Carroll RJ, Ruppert D, Stefanski LA. *Measurement error in nonlinear models.* Chapman & Hall: London, 1995.
29. Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods of information in medicine* 1990; **29**(1):12-22.
30. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agrees closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004; **57**:925-932.
31. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US--a meta-analysis. *Radiology* 2000; **216**(1):67-77.

32. Nederkoorn PJ, van der Graaf Y, Hunink MG. Duplex ultrasound and magnetic resonance angiography compared with digital subtraction angiography in carotid artery stenosis: a systematic review. *Stroke* 2003; **34**(5):1324-32.
33. Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG. MR imaging of the menisci and cruciate ligaments: a systematic review. *Radiology* 2003; **226**(3):837-48.
34. Harbord R, Steichen T. METAREG: Stata module to perform meta-analysis regression. In. revised 02 Feb 2005 ed: Boston College Department of Economics; 2004. p. Statistical Software Components S4446201.
35. van Houwelingen HC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2272-2284.
36. Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24):3497-3518.
37. McIntosh MW. The population risk as an explanatory variable in research syntheses of clinical trials. *Statistics in Medicine* 1996; **15**:1713-1728.
38. van Montfort K, Mooijaart A, de Leeuw J. Regression with errors in variables: estimators based on third order moments. *Statistica Neerlandica* 1987; **41**(4):223-237.
39. Chang B, Waternaux C, Lipsitz S. Meta-analysis of binary data: which within study variance estimate to use? *Statistics in Medicine* 2001; **20**(13):1947-1956.
40. Carlin B, Louis T. Bayes and Empirical Bayes Methods for Data Analysis. New York: Chapman & Hall; 1996.
41. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random effects meta-analyses: a comparative study. *Statistics in Medicine* 1995; **14**:2685-2699.
42. Kalaian HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. *Psychological Methods* 1996; **1**(3):227-235.
43. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**(22):2537-2550.

6

Meta-analysis of summary survival curve data

Abstract

The use of standard univariate fixed and random effects models in meta-analysis has become well known in the last twenty years. However these models are unsuitable for meta-analysis of clinical trials that present several survival estimates during a follow-up period illustrated by survival curves. Therefore special methods are needed to combine the survival curve data from different trials in a meta-analysis. For this purpose only fixed effects models have been suggested in the literature. In this chapter a multivariate random effects model is proposed for joint analysis of survival proportions reported at multiple times in different studies, to be combined in a meta-analysis. The model could be seen as a generalization of the fixed effects model of Dear, published in 1994. We illustrate the method by a simulated data example as well as a clinical data example of meta-analysis with aggregated survival curve data. All analyses can be done with standard general linear MIXED model software.

1 Introduction

Since the introduction in 1976[1] of the term 'meta-analysis', i.e. the quantitative approach to summarize the outcomes of several studies, it has become an increasingly important technique in clinical research. The two main statistical approaches to meta-analysis are the fixed effects model and the random effects model. Nowadays it is more or less common practice to analyse the data with the univariate random effects meta-analysis model as proposed by DerSimonian and Laird[2]. If one has more outcome parameters per study, usually each outcome is analysed separately. Despite of its many disadvantages also the fixed effects model is still used, thereby ignoring possible between-trial variation, leading to overestimation of the precision of the estimate and restricting the inference only to the trials included in the meta-analysis[3, 4].

With the increasing popularity of meta-analysis, also the field of application of meta-analysis is growing. In earlier days the main interest was to statistically pool the results of independent but 'combinable' studies[5] regarding one specific outcome variable at a time. This is still common practice in most meta-analyses. In the last years a new trend is recognizable in meta-analyses. Clinical interest does not concern only one specific outcome measure but the combination of several outcome measures that are presented in the individual studies. Especially the relationships between these outcome measures could be of special interest. A famous example is the many research that is done on the subject of the relationship between treatment effect and baseline risk[4, 6-16]. Another example is a meta-analysis of clinical trials with a relevant clinical outcome, called the 'true' endpoint, as well as an early response variable, called the 'surrogate' endpoint. The goal of such a meta-analysis is to investigate the association between treatment effects on the surrogate and true endpoint [17-22]. A clinical application of combining (three) multiple endpoints in a meta-analysis of clinical trials in which the univariate and multivariate approaches are compared, has recently been published[23].

Another important type of multiple endpoints arises when each study reports survival proportions at a series of time points. Naturally, within one study the reported survival proportions are correlated over time. This means that the data have a multivariate nature and the analysis becomes more complicated. This is the central issue of this chapter.

In the literature several methods are proposed for meta-analysis of survival data. If the meta-analysis concerns a comparison between two treatment groups, the simplest approach would be to summarize the difference between the two treatment arms of each contributing trial by a single number like the (log) hazard ratio, along with its

standard error, and use standard methods of meta-analysis to combine them[24, 25]. Whitehead and Whitehead[26] discussed the meta-analysis of survival data, combining efficient score statistics for the hazard ratio of an assumed proportional hazards model. However, they expressed doubt over the chances of finding enough information about the required statistics such as the log-rank test statistics, or the hazard ratio estimates and their standard errors. Parmar et al.[27] presented a number of methods of extracting estimates of these statistics in a variety of situations. But even with these suggestions, it is not always possible to calculate the necessary quantities. A minor problem of this approach might be that it cannot use information from single-arm trials [24].

The basic information that is reported on the survival in the two treatment groups is often just a series of survival proportions in a study for a number of time points like each year, or twice a year, but the choice of the follow-up times could easily be different across the studies. Therefore the data are mostly very unbalanced and difficult to analyse with standard methods. To tackle this problem, investigators often reduce the survival curve to one or some fixed points in time, e.g. the three-year survival rate. Then the data can be analysed with the standard univariate random or fixed effects model for each of the chosen time points separately [28]. If a trial does not report survival for a chosen time point, it can be left missing or an estimate can be imputed using inter- or extrapolation. Or, like in a recent article[29], follow-up data from years 2 to 3, 4 to 5, and 6 to 8 are combined and reported as three-, five- and eight-year end points, respectively. And for the subgroup of diabetic patients in this meta-analysis outcomes were reported 'where available' at 4 and 6.5 years. It is not clear what happens if these outcomes are not reported at these exact time points. Besides, the estimate of the 4-year end point could be based on completely other trials and patients than the estimate of the 6.5-year end point. In this approach, while the meta-analyst might be relieved to be able to use standard univariate statistical techniques, he completely neglects the information about the course of the survival rates over time.

Obviously, this last approach is not the optimal solution and better methods have been proposed[24, 30-33]. In these methods data concerning entire survival curves are combined in a meta-analysis, instead of artificially reducing the data to just one single survival statistic or to a survival estimate on just one fixed point in time. In the overview article of Earle et al.[34], all five methods found in the literature to combine published survival curves were assessed. In this assessment the resulting summary survival curve of each method was compared with the 'golden standard': the curve calculated from the corresponding individual patient data (IPD). A brief description of each of the five methods is given in the appendix. In this overview Earle et al.[34]

made clear that the most recent methods to combine aggregated survival curve data were published in 1994 and that all the methods are fixed effects models. Although the method of Dear[24] was not significantly more accurate than the other models, it is one of the most recommended methods according to the review of Earle et al.[34].
The model of Dear[24] is an extension of the method of Raudenbush et al.[35], who showed how to analyze effect sizes for two or more outcomes jointly in a fixed-effects generalized-least-squares (GLS) regression model that allows adjustment for study-level covariates[36]. Dear[24] showed how to estimate the correlations among the serial survival proportions, allowing the survival proportions reported at multiple times by the trials to be analysed together in a fixed-effects model[36]. Berkey et al.[36] demonstrated that fixed-effect regression models for correlated outcomes may seriously underestimate the standard errors of regression coefficients when the regression model does not explain all the among-trial heterogeneity. Therefore Berkey et al.[36] proposed a random-effects approach for the regression meta-analysis of multiple correlated outcomes. In the Tutorial of van Houwelingen et al.[4] is shown how this model can be fitted easily with standard software.
In this chapter we propose a multivariate random effects model for joint analysis of survival proportions reported at multiple times in different studies. The method makes use of the complete and possibly unbalanced set of reported survival proportions in all studies, and no inter- or extrapolation to common chosen endpoints is needed. The model could be seen as a generalization of the fixed effects model of Dear[24] or as a combination of the models of Dear[24] and Berkey et al.[36] and is fitted with standard software as described in van Houwelingen et al.[4]. The method is applied on a simulated data example as well as on a clinical data example of meta-analysis with aggregated survival curve data, which are described in Section 2. The clinical data example has also been used by Dear to illustrate his method. In Section 3 the model of Dear[24] is presented, after which we propose our generalization of that model. In Section 4 the results of our model are discussed for the simulated data example, and the results of the clinical data example are compared to the results of Dear[24]. In Section 5 we conclude with a discussion.

2 Data sets

To illustrate our method, we use two datasets. The first one is a simulated dataset, because of the lack of raw, unbalanced, survival curve meta-analysis data sets in the literature. This data set is described in Section 2.1. The second data set is a real clinical data example and is described in Section 2.2. Since Dear used the same clinical data

set, we will compare the results of our method to the results of the method of Dear later on.

2.1 Simulated dataset

In real life the world of a meta-analyst can be quite complicated. A number of scientific papers relevant for the research question can be found, but in most of the cases the authors publish very different kinds of results. To get an example, we made a meta-analysis data set consisting of 10 trials, each containing a treatment arm and a control arm. In this example the first author presents survival rates after each half a year during three years. The second author only presents survival rates at whole years, that means after 1, 2 and 3 years. The third author publishes annually survival rates for each half a year, so after 0.5, 1.5 and 2.5 years. In the trial of the fourth author the survival rates are reported after one year and after three years, while the fifth author gives the survival rates after one and two years. The sixth author presents for the control group only the estimates at one and three years, but gave an extra survival estimate for the experimental group at two years. The seventh author presents the experimental group almost similar to the first author, i.e. every half year during 2.5 years. However, for the control group this author gave the estimates only at 0.5 year and at 2.5 years. The eighth author just reported survival proportions for both treatment groups at 1 and 3 years, while the ninth author did a strange thing and reported the control group only at baseline (1 year) and the experimental group only after 3 years. Finally the tenth author gives for both groups simply one survival estimate after 2.7 years. Nobody else gave survival rates after 2.7 years. See Table 1 for the data.

Unfortunately we could not find a published meta-analysis of survival time data in the literature with such a realistic but chaotic data structure. Many meta-analysts who are confronted with so many structural missing data in the dataset, will choose one, two or maybe three fixed time points and meta-analyse the survival rates at the different chosen time points separately from the survival rates at other time points. Also some meta-analysts might get confused and will not know what to do with the survival rates measured at half a years. In most of the cases they will ignore that information or they will inter- and extrapolate the measurements of e.g. trial 3 to get survival rates at 1, 2 and 3 years. And what to do with trial 10 with measurements at 2.7 years? The most easy and logical choice would be to estimate the 2 or 3 years survival for that trial, but this would be pretty hard because there is only one survival measurement per treatment arm in that trial. Sometimes the meta-analyst will state that the 2.7 years survival rate is set to a time point of 3 years, but of course this kind of decision is very arbitrary.

Table 1. Data from the simulated data set: Survival (standard errors in parentheses) of trial *i* by treatment arm j and year k.

Time:	**k = 0.5**	**k = 1.0**	**k = 1.5**	**k = 2.0**	**k = 2.5**	**k = 2.7**	**k = 3.0**
Trial *i*:							
1 control (j=1)	0.74 (0.04)	0.58 (0.05)	0.52 (0.05)	0.50 (0.05)	0.47 (0.05)	.	0.43 (0.05)
1 exp (j=2)	0.80 (0.04)	0.74 (0.04)	0.66 (0.05)	0.62 (0.05)	0.52 (0.05)	.	0.48 (0.05)
2 control (j=1)	.	0.50 (0.05)	.	0.25 (0.04)	.	.	0.16 (0.04)
2 exp (j=2)	.	0.67 (0.05)	.	0.45 (0.05)	.	.	0.37 (0.05)
3 control (j=1)	0.77 (0.04)	.	0.52 (0.05)	.	0.29 (0.05)	.	.
3 exp (j=2)	0.80 (0.04)	.	0.53 (0.05)	.	0.40 (0.05)	.	.
4 control (j=1)	.	0.51 (0.05)	.	.	.	.	0.09 (0.03)
4 exp (j=2)	.	0.68 (0.05)	.	.	.	.	0.31 (0.05)
5 control (j=1)	.	0.63 (0.05)	.	0.45 (0.05)	.	.	.
5 exp (j=2)	.	0.91 (0.03)	.	0.71 (0.05)	.	.	.
6 control (j=1)	.	0.57 (0.05)	.	.	.	.	0.23 (0.04)
6 exp (j=2)	.	0.78 (0.04)	.	0.52 (0.05)	.	.	0.42 (0.05)
7 control (j=1)	0.78 (0.04)	.	.	.	0.12 (0.03)	.	
7 exp (j=2)	0.88 (0.04)	0.58 (0.05)	0.43 (0.05)	0.28 (0.05)	0.19 (0.04)	.	
8 control (j=1)	.	0.69 (0.05)	.	.	.	.	0.23 (0.04)
8 exp (j=2)	.	0.78 (0.04)	.	.	.	.	0.47 (0.05)
9 control (j=1)	.	0.68 (0.05)	.	.	.	.	.
9 exp (j=2)	.	.	.	.	.	.	0.42 (0.05)
10 control (j=1)	.	.	.	.	.	0.19 (0.04)	.
10 exp (j=2)	.	.	.	.	.	0.47 (0.05)	.

The best way would be to analyse all available data simultaneously, irrespective of the multiple time points on which the survival rates are measured. The approach proposed in this chapter provides a framework in which the kind of data presented in Table 1 can be modelled.

To simulate the data we assumed Weibull distributed survival times. The cumulative survival functions of the treatment groups in trial i were $\exp\left(-\exp(\beta_{0i}+\beta_{1i}Z)t^{\alpha_i}\right)$, with α_i the shape parameter and $\exp(\beta_{0i}+\beta_{1i}Z)$ the scale parameter, where Z is the treatment indicator. For the three parameters we choose the following independent normal distributions:

α_i ~ N(1.0 , 0.04)

β_{0i} ~ N(-0.7 , 0.04)

β_{1i} ~ N(-0.5 , 0.04)

It follows that on the log minus log survival scale the relation with log time is linear: $\ln(-\ln(S(t))) = \beta_0 + \beta_1 Z + \alpha \ln(t)$. We randomly drew 100 survival times per treatment group.

To introduce censoring, we assumed a research project with a total study period of four years. The intake period of the project was two years, followed by a follow-up period of another two years. Thus the potential follow-up period per patient varied between two and four years. For each patient we randomly draw a censoring time from a uniform distribution on the interval from 2 to 4 years. When this censoring time was lower than the survival time of that patient, the patient was censored. In our data this resulted in 30% censored patients. Next, to get the survival estimates for the different time points together with their standard errors, we used Kaplan-Meier survival analysis.

2.2 Clinical data example: Bone-marrow transplantation versus Chemotherapy

A drawback of simulated data is that the data might not be realistic. To illustrate the proposed method with real life data, we use the meta-analysis data provided by Begg et al.[37]. The same data were used by Dear[24] to illustrate his method. In that meta-analysis 14 studies were included in which alternative therapeutic approaches were followed to treat acute non-lymphocytic leukaemia in young adults. The authors studied the relative efficacy of bone-marrow transplantation (BMT) versus conventional chemotherapy for patients in first complete remission. In six studies, of which four randomised, the two treatment arms were directly compared to each other. Eight observational studies included patients on only one of the two treatments.

The data consist of the Kaplan-Meier probabilities of disease-free survival at a maximum of five 1-year intervals after start of treatment together with their standard errors. Individual patient data are not available. The complete dataset is shown in Table 2. Every line in Table 2 corresponds to one clinical trial, the first six studies include both treatment arms, two of the studies include only patients on BMT and six studies include only patients on chemotherapy. All studies give estimates for at least 3 years after start of treatment. The raw survival curves corresponding to the data in Table 2 are illustrated in Figure 1.

Table 2. Data from Begg et al. (1989): Percent disease-free survival (standard errors in parentheses) of trial *i* by treatment arm j and year k (1 to 5)

Trial (i)	**BMT (j=1)**					**Chemotherapy (j=2)**				
	k=1	k=2	k=3	k=4	k=5	k=1	k=2	k=3	k=4	k=5
1	49 (12)	46 (12)	42 (12)	40 (12)	40 (12)	54 (8)	25 (8)	23 (7)	23 (7)	23 (7)
2	55 (10)	50 (10)	36 (9)			40 (8)	23 (7)	23 (7)	23 (7)	
3	54 (10)	47 (13)	40 (13)	40 (13)		54 (9)	42 (8)	28 (8)	28 (8)	
4	70 (23)	70 (23)	70 (23)	70 (23)		48 (17)	48 (17)	17 (13)		
5	54 (4)	46 (5)	42 (6)			40 (5)	21 (4)	16 (4)	16 (4)	
6	54 (2)	43 (3)	40 (3)	39 (3)		50 (4)	32 (4)	24 (4)	18 (4)	
7	59 (8)	49 (9)	47 (9)	47 (9)	47 (9)					
8	61 (8)	53 (8)	53 (8)	53 (8)	53 (8)					
9						60 (9)	48 (9)	32 (9)	32 (9)	32 (9)
10						44 (5)	26 (4)	17 (5)	16 (4)	
11						50 (3)	33 (3)	26 (3)	22 (3)	19 (3)
12						62 (3)	38 (3)	29 (3)	24 (3)	22 (3)
13						50 (10)	24 (8)	16 (7)	12 (6)	
14						76 (7)	53 (8)	53 (8)	50 (8)	50 (8)

Like many other authors with similar data [29, 38], Begg et al.[37] analysed the data separately for each year. As we already mentioned, this is the simplest way to carry out meta-analysis of survival data, because one can use standard univariate meta-analysis methods[24, 25]. However, doing separate analyses for each point in time and thus carrying out many meta-analyses, is inefficient and could lead to inappropriate conclusions[27]. It can lead to loss of power, because in each analysis only a portion of the data is used.

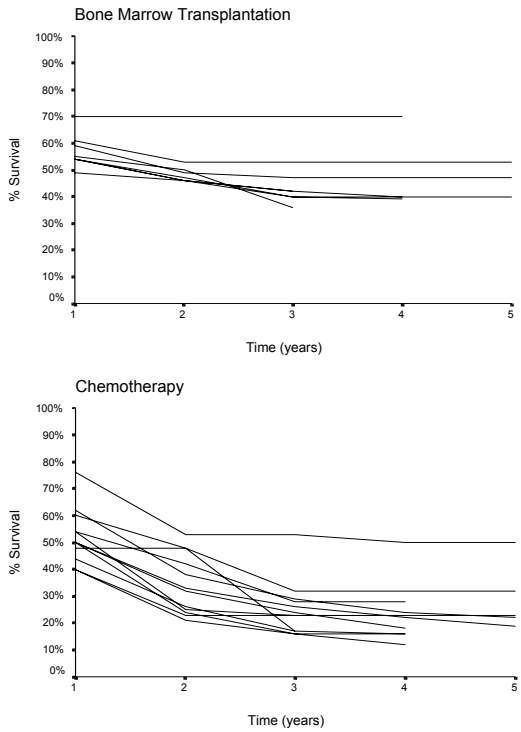


Figure 1. Survival curves based on data from Begg et al. (1989).

Next, it can give rise to a multiple testing problem and it could be difficult to interpret the results. Furthermore, this approach is only sensible if the times for which survival estimates are available are the same across studies. Finally, the results may be biased due to selective missing data, as might be the case for k=4 and 5 in our example.

3 Methods

The studies to be combined in the meta-analysis are indexed by i, in our clinical data example $i = 1,...,14$. In each study one or more treatments are considered. For each of the studies survival estimates are available for treatment j, where in our clinical data example j=1 for BMT and j=2 for chemotherapy. Some of the studies have survival estimates for both j=1 and j=2, and some of the studies –the observational ones– only have survival estimates for either j=1 or j=2. For each treatment arm at times t_{ijk}, where the index k counts the time points (at most 5 in the clinical data example), a survival

estimate is available together with its standard error. The true survival probability of the j^{th} treatment in the i^{th} trial on time point t_{ijk} is denoted by s_{ijk}. The estimate of s_{ijk} is denoted by $\hat{s}_{ijk}$. The corresponding standard error is denoted by se_{ijk}. The estimated correlations between the survival estimates would preferably also be available, but unfortunately this will seldom be the case. In general, estimates of correlations between multiple outcome measures often come from some external source[39]. However, in the special case of estimated survival probabilities, the correlations can be estimated from the data, as is shown by Dear[24].

As shown in the simulated data example, the pattern of the time points might be different across the trials and within the trials across the treatment groups, dependent on the choice of the authors of the time points for which to provide survival estimates and standard errors in their publications. So, although in our clinical data example the time points are fixed to years after start of treatment, this is not necessarily the case.

3.1 The method of Dear

As a stepping stone to the method we propose in this chapter, we briefly discuss the generalized least squares method as proposed by Dear in 1994[24]. In this approach a generalized linear regression model is used to relate the estimated survival proportions $\hat{s}_{ijk}$ to a design matrix X with both between- and within-study covariates such as time, study and treatment characteristics, including interaction terms. Dear treated all covariates as categorical represented by dummies in the model, but that is not necessary. The model is:

$$\hat{s}_i = X_i\beta + \varepsilon_i \qquad (1)$$

where $\hat{s}_i$ is the column vector of $\hat{s}_{ijk}$'s and ε_i a column vector of residuals with $\varepsilon_i \sim \mathrm{N}(0, V)$

The ε's are assumed to be independent between studies and treatment arms. Since the errors ε_{ijk} of serial observations in the same treatment arm within the same study are bound to be related, the off-diagonal elements of matrix V will not all be zero. V is block diagonal with blocks corresponding to the treatment groups within studies. The main diagonal is set equal to the reported squared standard errors se^2_{ijk} of the survival proportions (Table 2). To estimate the covariances within a treatment group, Dear[24] made use of the fact that the correlations of proportions between time points t_{ijk} and $t_{ijk'}$ are given by

$$\mathrm{corr}(s_{ijk}, s_{ijk'}) = \sqrt{\frac{s_{ijk}(1-s_{ijk'})}{(1-s_{ijk})s_{ijk'}}}$$

This formula is derived by exploiting the fact that the estimated cumulative hazards over different intervals are independent. Based on these correlations, the covariances in matrix V are obtained by multiplying the correlation with the corresponding standard errors given in Table 2.

$$\mathrm{cov}(\hat{s}_{ijk}, \hat{s}_{ijk'}) = se_{ijk} \cdot \sqrt{\frac{s_{ijk}(1-s_{ijk'})}{(1-s_{ijk})s_{ijk'}}} \cdot se_{ijk'} \qquad (2)$$

The se_{ijk}'s are considered fixed and known, while the s_{ijk}'s are to be estimated. The β parameters are estimated in an iterative manner. Iterations start with substituting the observed survival proportions in (2). Then, given this covariance matrix V, the β's are estimated with generalized least squares (GLS): $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}s$. With these β's new estimates of the s_{ijk}'s are computed by (1) and substituted in (2). This results in a new matrix V, which subsequently is used to calculate new GLS β estimates. This is repeated until convergence, providing fully efficient maximum likelihood estimators for the β parameters.

Compared with the other methods published in the literature (see Appendix), the model of Dear[24] is very generally applicable. With the model of Dear[24] one can do a joint analysis of survival proportions at multiple times, instead of analysing the survival probabilities separately for each time interval. Also, it is possible to combine studies with a different number of curves, as in our example where some comparative studies include two treatment arms and some observational studies have only one treatment arm in the study. Actually this only works under the assumption that the same treatment-specific profile of survival over time applied in all studies. When this assumption is doubtful, then study characteristics, such as patient population parameters, should be sought to account for the discrepancies. These characteristics can be represented in the model through study-level covariates[24]. And finally, with the model of Dear it is possible to fit and compare different regression models, which makes the model very flexible and informative.
However, the model of Dear has one important shortcoming in that it is a fixed effects model. To allow for between study heterogeneity, Dear introduced a dummy variable for each study, indicating the common survival level of each study. A disadvantage of this is that it results in very many parameters in the model relative to the number of data points. Moreover the inference is restricted to the studies at hand instead of 'all similar trials'.

There are also some disadvantages of the model, which are not true shortcomings of the model, but refer to the way in which this method is usually applied in practice. Dear presented his model in such a way that one needs a fixed pattern of time points

for which survival estimates are available, like every year in our example. Usually this kind of balanced data will not be provided in the publications of the studies and one needs to intra- and extrapolate to get estimates on exactly the same time points across the studies. Next, Dear used dummies for all covariates including time and study, which can result in an enormous amount of parameters to be estimated, especially when one also has to include also some interaction terms. And finally, Dear fitted a linear model on the survival probabilities, but of course the probabilities are restricted to the interval from 0 to 1. Therefore, a linear model could give non-sensible fitted values. However, time could also be modelled as a continuous covariate, enabling different time patterns between studies and reducing the number of parameters. Also, the method could be applied to transformed survival probabilities, e.g. with the logit or log(-log) transformation.

3.2 Multivariate random effects model

In this chapter we propose a multivariate random mixed effects model that relates the log-minus-log transformed survival estimates ln(-ln($\hat{s}_{ijk}$)) to both fixed and random covariates, like time or ln(time), treatment group etc. Any other transformation of the survival probabilities that maps the interval [0,1] into (-∞, ∞), for instance the logit transformation, might be chosen as well, but we prefer the ln(-ln) for reasons to be discussed later on. We assume the following model:

$$\ln(-\ln(\hat{s}_i)) = X_i\beta + Z_i b_i + \varepsilon_i \qquad (3)$$

with
$b_i \sim \mathrm{N}(0, D)$,
$\varepsilon_i \sim \mathrm{N}(0, V_i)$

and

$$V_i = \begin{pmatrix} V_{i1} & 0 \\ 0 & V_{i2} \end{pmatrix} \text{ with } \quad V_{ij} = \frac{se_{ijk}}{\hat{s}_{ijk}\ln(\hat{s}_{ijk})}\sqrt{\frac{s_{ijk}(1-s_{ijk'})}{(1-s_{ijk})s_{ijk'}}}\frac{se_{ijk'}}{\hat{s}_{ijk'}\ln(\hat{s}_{ijk'})} \qquad (4)$$

In equation (3), $\hat{s}_i$, X_i and ε_i have the same meaning as earlier. β is the parameter vector containing the fixed effects (time, treatment, etc.). Compared with Dear's model, the model is extended with the random part $Z_i b_i$. The vectors of random coefficients b_i are assumed to be independent normally distributed with expectation zero and between studies covariance matrix D, independent from the ε_i's. Z_i is the design matrix for the random effects, typically containing intercept, time and possibly treatment effect. The residual components have expectation zero and covariance matrix V_i, which is in fact the within-trial covariance matrix. Since the residual components across time are correlated within a trial arm (or survival curve) but

independent between treatment arms, the covariance matrix V_i is a block diagonal matrix existing of blocks corresponding to the treatment arms. In our clinical example we get two matrices V_{i1} (for the MBT survival curves) and V_{i2} (for the Chemotherapy survival curves). The within study covariance matrix (4) is completely analogous to (2). Notice that the standard error of the log(-log) transformed observed survival probability is:

$$\log(-\log(\hat{s}_{ijk})) = \frac{se_{ijk}}{\hat{s}_{ijk}\ln(\hat{s}_{ijk})}$$

It is assumed to be known. The correlation between two transformed survival estimates is equal to

$$corr(\log(-\log\hat{s}_{ijk}), \log(-\log\hat{s}_{ijk'})) = \sqrt{\frac{s_{ijk}(1-s_{ijk'})}{(1-s_{ijk})s_{ijk'}}}$$

Analogous to Dear's approach, this correlation is estimated from the data.

The parameters in the model, the β's and the between studies covariance matrix D, are estimated similar to Dear's approach in an iterative fashion as follows. In the first step, an initial estimate of V_i is obtained by substituting the observed survival probabilities into (4). With this covariance matrix the mixed model is fitted and the new survival estimates are used to calculate the new correlations in (4) and so on. This can be done easily in a General Linear Mixed Model program provided that the residual variances can be fixed at arbitrary values per individual survival estimate[4]. Because of the iterative manner of model fitting, it would also be convenient if the fitted survival estimates could be saved in order to automatically update the correlations between them and thereby the covariance matrix V_i. These features are for instance available in the procedure MIXED of SAS and the function *lme* of S-Plus, but it might also be available in other statistical software.

By applying a transformation to the observed survival probabilities, our model guarantees that fitted survival probabilities are between 0 and 1. However, as for Dear's method, the fitted survival curves are not necessarily non-increasing. We think that in practice non-monotonically non-increasing fitted curves will be very rare and therefore will not be a serious problem. Also, if curves are extrapolated from the smallest t_{ijk} to t=0, the fitted survival at t=0 might be smaller than 1.

As an illustration we apply this method to our two data examples and we compare the results of the clinical data example with the results given in the publication of Dear[24], who used the same dataset.

4 Results

4.1 Results simulated dataset

Using Proc MIXED of SAS, we fitted the following model on the data of Table 1.

$$\log(-\log \hat{s}_{ijk}) = \beta_0 + \beta_1 \text{treat}_{ijk} + \beta_2 \log(\text{year}_{ijk}) + b_0 + b_1 \text{treat}_{ijk} + b_2 \log(\text{year}_{ijk}) + \varepsilon_{ijk} \quad (4)$$

Here *treat* is a dummy variable for the treatment, 0=control and 1=experimental. We allow random effects for intercept and slope of log time and a random treatment effect. We assume a zero mean multivariate normal distribution for the random effects (b_1,b_2,b_3) with a covariance matrix which is completely left free and has to be estimated. Choosing the log(-log) transformation for the survival proportions and log(year) instead of year itself as covariate, corresponds to a Weibull distribution assumption. The advantage is that the treatment effect β_2 is expressed as a hazard ratio. Furthermore, survival curves start at level 1 at t=0. Of course, in practice the Weibull assumption might not be true, and other covariate specifications could be tried. Also another transformation of the survival probabilities than the log(-log), such as the logit of probit, might be entertained. The parameters are estimated by repeated calls of Proc MIXED, each time updating the correlations. The results are given in Table 3.

Table 3. Results of fitting model (4) on the data of Table 1.

regression coefficients	**estimate**	**standard error**
intercept	-0.6143	0.0698
ln(year)	0.9705	0.0697
treat	-0.4970	0.0858
covariance parameters		
variance *intercept*	0.0243	
covariance intercept* ln(year)	0.0090	
variance *ln(year)*	0.0311	
covariance intercept* treat	0.0098	
covariance ln(year)*treat	-0.0144	
variance *treat*	0.0322	

The overall mean estimated survival curves of both treatment groups together with their confidence intervals are drawn in Figure 2.

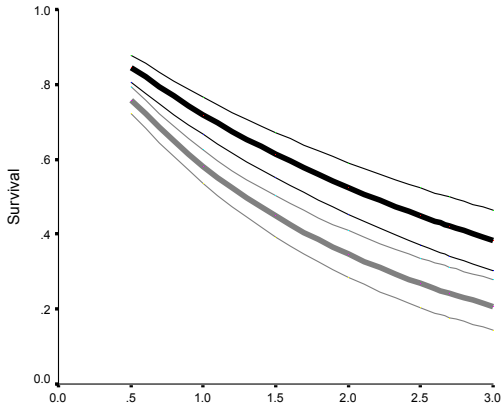


Figure 2. Overall mean survival curves (plus confidence bands) of the treated group (black lines) and the control group (grey lines) estimated from the data in Table 1.

All between study variances differed significantly from zero when tested with the likelihood ratio test. As a model check, it was investigated whether adding terms as $\ln(year)^2$, interaction between ln(year) and treatment did improve the model, but no extension was statistically significant.
As a by-product of the analysis, empirical Bayes estimates are provided for the study specific survival curves. These are illustrated in Figure 3.

4.2 Results clinical data example

In this section we present the results of the GLS-model of Dear[24] as well as our multivariate random effects model to see the differences and the similarities.
Dear fitted a linear model to relate the estimated survival proportions to several dummy variables, indicating the treatment, year of follow-up and the study. The final model in the publication of Dear includes dummy variables for 'study', 'treatment', 'follow-up year' and for the interaction of the dummies 'treatment by follow-up year', resulting in the following model:

$$\hat{s}_{ijk} = \beta_0 + \sum_{i=1}^{14} \beta_i study_i + \sum_{j=1}^{2} \beta_j treatment_j + \sum_{k=1}^{5} \beta_k year_k + \sum_{j=1,k=1}^{j=2,k=5} \beta_{jk} (treatment * year)_{jk} \quad (5)$$

with constraints $\sum \beta_i = 0, \sum \beta_j = 0, \sum \beta_k = 0$ to ensure estimability. The covariance matrix of $\hat{s}_{ijk}$ is estimated using generalized least squares in an iterative way, as explained in Section 3 of this chapter.

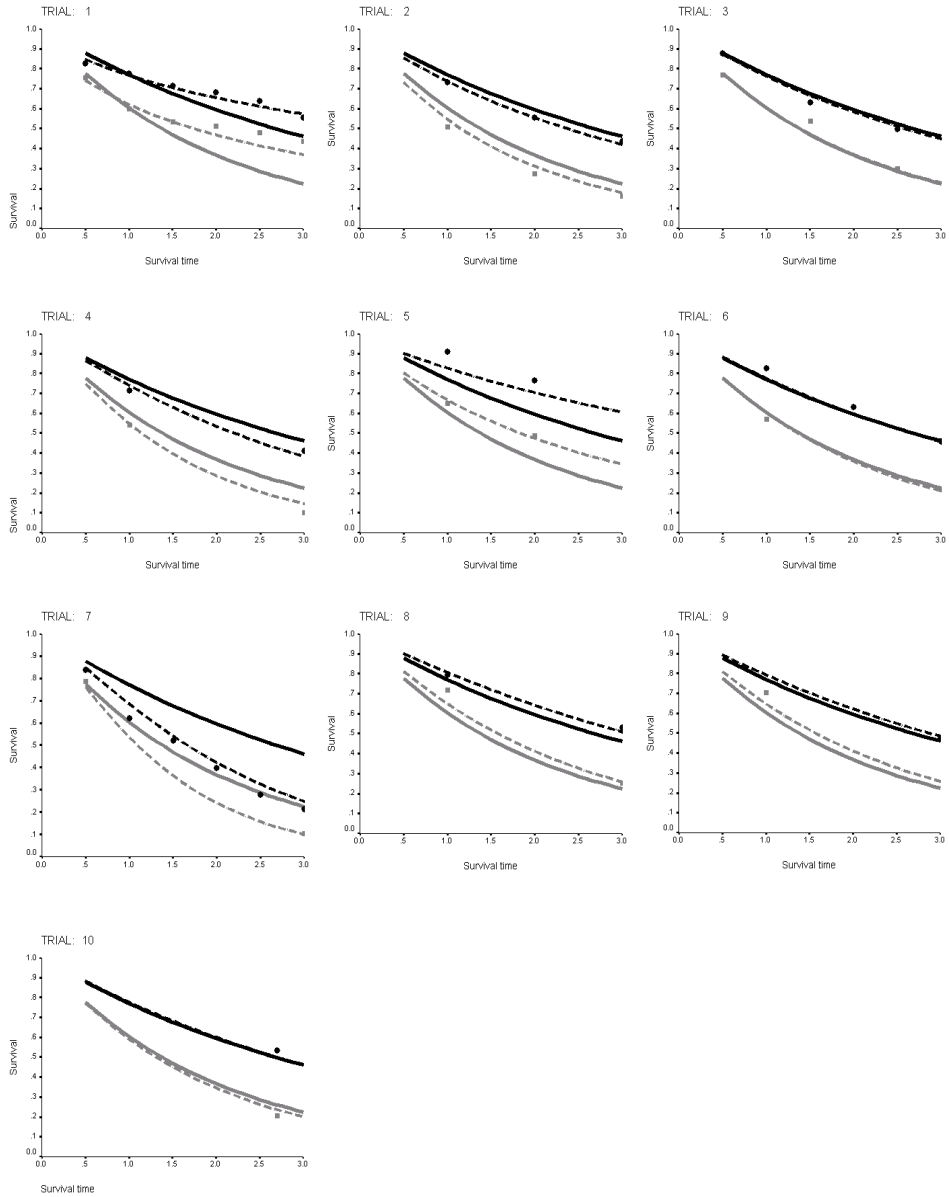


Figure 3. Estimated study specific (empirical Bayes) survival curve lines, together with the estimated mean survival curves per treatment (black = treated, grey = control).

To compare this model to our multivariate random effects model, we started to include the same variables as Dear did, but in a continuous way. So, with estimated log-minus-log survival as dependent variable, treatment (0=BMT, 1=Chemotherapy), ln(year), and the interaction between ln(year) and treatment were included as covariates, along with a random intercept and regression coefficients for ln(year) and treatment. Adding ln(year)^2 and the interaction between ln(year)^2 and treatment significantly improved the model. The random terms for ln(year) and treatment turned out to be non-significant and were dropped from the model. Thus we ended up with the following model

$$\log(-\log(\hat{s}_{ijk})) = \beta_0 + \beta_1 treat + \beta_2 \ln(year) + \beta_3 treat * \ln(year) + \beta_4 \ln(year)^2 + \beta_5 \ln(year)^2 * treat + b_{0i} + \varepsilon_{ijk} \quad (6)$$

The results are given in Table 4.

Table 4. Parameter estimates of model (6)

	beta (se)	**p-value**	**var(b_{0i})**
intercept	-0.61 (0.08)	<0.0001	0.04
treatment	0.12 (0.08)	0.11	
ln(year)	0.42 (0.07)	<0.0001	
ln(year)^2	-0.10 (0.04)	0.02	
treatment*lnyear	0.54 (0.09)	<0.0001	
treatment*lnyear^2	-0.14 (0.05)	0.01	

In Figure 4 the mean survival estimates per treatment are shown for models (5) and (6), together with their confidence intervals.

The mean survival curves and confidence intervals are quite similar in both models. Notice however that the model of Dear included 23 β parameters in the model (to estimate 85 survival probabilities) versus only 7 (including only the variance of the intercept) in our model.

The empirical Bayes study specific survival curves are depicted in Figure 5. The shrinkage phenomenon is nicely illustrated in this figure. E.g. trial 4 has a much higher observed survival curve for the BMT treatment compared to the other trials (see Figure 2 and Table 2). However, it is a very small trial with large standard errors (Table 2). So, the empirical Bayes estimate of the survival curve for this trial has strongly shrunk towards the average survival curve.

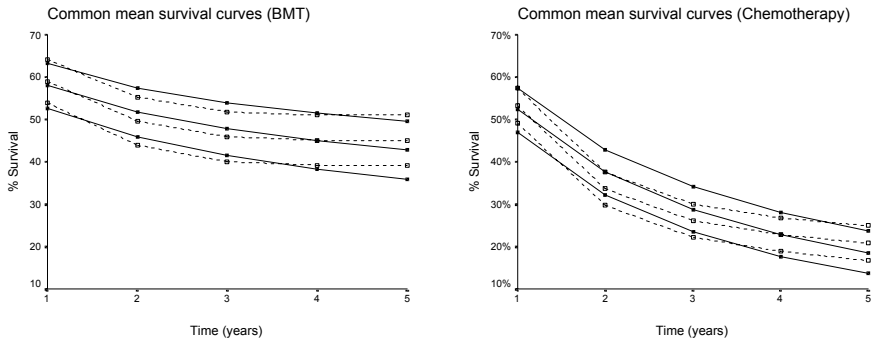


Figure 4. Mean survival estimates per treatment, with 95% confidence limits. The dotted lines represent Dear's model, the solid lines represent the multivariate random effects model.

On the other hand, also trial 14 has an extremely high survival curve for Chemotherapy compared to the other trials. However, since trial 14 is a large trial with relatively small standard errors, the empirical Bayes estimate of this survival curve has somewhat shrunk towards the common mean, but is remains on a higher level than the other survival curves.

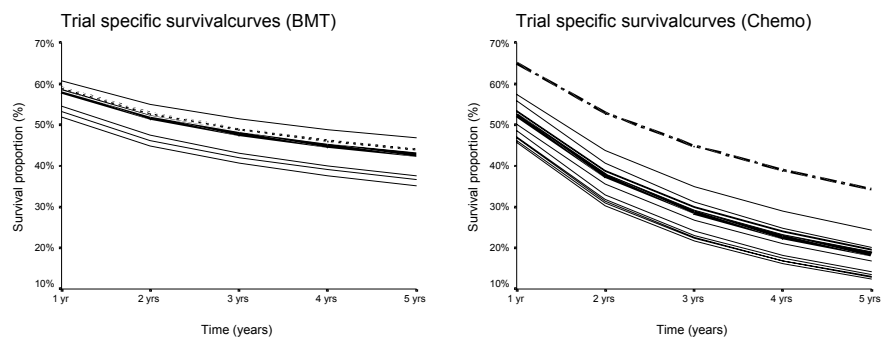


Figure 5. Trial specific survival curves of both treatment arms. The dashed line in the BMT plot represents the survival curve of trial 4. The dashed line in the Chemotherapy plot represents trial 14.

5 Discussion

In 1994 Dear proposed a general linear model with the survival estimate as dependent variable, and follow-up time, treatment and study as categorical covariates. The parameters are estimated by weighted generalised least squares in an iterative way. The main drawback of the model is that it is a fixed effects model. In this chapter we generalized the GLS-method of Dear[24] to a multivariate random effects framework like for instance proposed by Berkey et al.[36]. The model can be fitted in standard programs like SAS Proc MIXED. The method can also be considered as a generalisation of the DerSimonian-Laird random effects model for univariate outcomes[2]. For a fixed time *t*, our model reduces to the DerSimonian-Laird model. The modelling approach is very flexible in that the data set does not need to be balanced. Different studies may provide different numbers of survival estimates at different times. This enables the meta-analyst to analyse all available data as provided in the publications, without need to inter- or extrapolate to fixed times. There is also a lot of freedom in the modelling process. For instance, other transformation than the log minus log might be chosen, the shape of the survival curves could be modelled using regression splines or fractional polynomials etc, while still standard programs can be used. Our preference in first instance is the log minus log transformation in combination with ln(time) as covariate, since than the covariate effects can be interpreted as hazard ratio's as in a Cox regression model.

Similar as for Dears method, a disadvantage of our approach is that the estimated curves are not forced to be survival curves. They are not necessarily non-increasing, and when extrapolated to 0, the curve does not necessarily start at 1. We do not think that this is a serious disadvantage in practice, but there is certainly a need for models in which the curves in a natural way are forced to be real survival curves.

We fitted our models by iterated linear mixed model fits, updating the estimated correlations between survival estimates of the same curve. An open question is whether the empirical Bayes survival estimates or the estimates based on only the fixed part should be used. Also it might be possible to update the standard errors of the transformed survival estimates as well. In our examples these alternatives gave similar results. Simulation studies might be carried out to what the best method is.

References

1. Glass, GV. Primary, secondary and meta-analysis of research. *Educational Researcher* 1976; **5**:3-8.
2. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177-188.
3. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicne* 1996; **15**:619-629.
4. van Houwelingen HC, Arends LR, Stijnen T. Tutorial in Biostatistics: Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589-624.
5. Egger M, Ebrahim S, Smith GD. Editorial: Where now for meta-analysis? *International Journal of Epidemiology* 2002; **31**:1-5.
6. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* 1992; **11**(16):2077-2082.
7. van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**(24):2273-2284.
8. McIntosh, MW. The population risk as an explanatory variable in research syntheses of clinical trials. *Statistics in Medicine* 1996; **15**:1713-1728.
9. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735-738.
10. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741-2758.
11. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2883-2900.
12. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine* 1998; **17**(17):1923-42.
13. van Houwelingen H, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1999; **18**(1):110-115.
14. Bernsen RMD, Tasche MJA, Nagelkerke NJD. Some notes on baseline risk and heterogeneity in meta-analysis. *Statistics in Medicine* 1999; **18**(2):233-238.
15. Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Statistics in Medicine* 2000; **19**(23):3251-3274.

16. Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24): 3497-3518.
17. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**(17):1965-1982.
18. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; **1**(1):49-67.
19. Gail MH, Pfeiffer R, Van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**(3):231-246.
20. De Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. *Controlled clinical trials* 2001; **22**(5):485-502.
21. Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled clinical trials* 2002; **23**(6):607-625.
22. Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal,* 2002; **44**(8):921-935.
23. Arends, L.R., Z. Vokó, and T. Stijnen, *Combining multiple outcome measures in a meta-analysis: an application.* Statistics in Medicine, 2003; **22**: p. 1335-1353.
24. Dear KBG. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994; **50**:989-1002.
25. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for meta-analysis in medical research*. Wiley: Chichester, 2000.
26. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991; **10**:1665-1677.
27. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* 1998; **17**:2815-2834.
28. Vokó Z. *Etiology and prevention of stroke. The Rotterdam Study.* Erasmus University Rotterdam, Department of Epidemiology and Biostatistics: Rotterdam, 2000.
29. Hoffman SN, TenBrook JA, Wolf MP, Pauker SG, Salem DN, Wong JB. A meta-analysis of randomized controlled trials comparing coronary artery bypass graft with percuteous transluminal coronary angioplasty: one and eight-year outcomes. *Journal of the American College of Cardiology* 2003; **41**:1293-1304.

30. Hunink MGM, Wong JB. Meta-analysis of failure-time data with adjustment for covariates. *Medical Decision Making* 1994; **14**:59-70.
31. Shore T, Nelson N, Weinerman B. A meta-analysis of stages I and II Hodgkin's disease. *Cancer* 1990; **65**:1155-1160.
32. Voest EE, van Houwelingen JC, Neijt JP. A meta-analysis of prognostic factors in advanced ovarian cancer with median survival and overall survival (measured with the log(relative risk)) as main objective. *European Journal Cancer Clinical Oncology* 1989; **25**:711-720.
33. Reimold SC, Chalmers TC, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic arterial fibrillation: observations on the role of trial design and implications of drug-related mortality. *American Heart Journal* 1992; **124**:924-932.
34. Earle CC, Wells GA. An assessment of methods to combine published survival curves. *Medical Decision Making* 2000; **20**:104-111.
35. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychological Bulletin* 1988; **103**(1):111-120.
36. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537-2550.
37. Begg C, Pilote L, McGlave P. Bone marrow transplantation versus chemotherapy in acute non-lymphocytic leukemia: A meta-analytic review. *European Journal of Cancer and Clinical Oncology* 1989; **25**:1519-1523.
38. Vale CL, Tierney JF, Stewart LA. Effects of adjusting for censoring on meta-analyses of time-to-event outcomes. *International Journal of Epidemiology* 2002; **31**:107-111.
39. Berkey CS, Anderson JJ, Hoaglin DC. Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine* 1996; **15**:537-557.
40. Chang B, Waternaux C, Lipsitz S. Meta-analysis of binary data: which within study variance estimate to use? *Statistics in Medicine* 2001; **20**:1947-1956.

Appendix

Statistical method	*Description*
Iterative generalized least squares (IGLS)[24]	A generalized linear regression model to relate survival proportions to between-trial and within-trial covariates. The correlation structure between successive survival proportions is derived iteratively from their fitted values.
Meta-analysis of failure-time data (MFD)[30]	Combination of failure-time data from various cohort studies, adjusting for differences in case-mix among studies by use of covariates. The model is based on the proportional-hazards model and the actuarial life-table approach.
Nonlinear regression (NLR)[31]	A survival curve in the form of an exponential decay function for nonlinear regression, with contributions of individual studies at each time point being the covariates.
Log(Relative Risk) (LRR)[32]	The survival function of each study is transformed using the log(-log)function which gives the log(relative risk), from which the average log(relative risk) curve is computed. The inverse transformation gives the summary survival curve.
Weighted LRR (w-LRR)[33]	This method does the same as the LRR-model, but weights the log relative risks with their inverse variances.

7

Discussion

Discussion

Since meta-analysis became popular in medicine about 25 years ago, biostatisticians have been interested and involved in developing statistical meta-analytic methods. In the beginning attention was focused on methods for deriving a common effect across studies. Later on, the focus moved to quantifying and reporting the heterogeneity between studies. This shift in focus parallels the transition in popularity from fixed effects models to random effects models[1]. Nowadays the random effects model, that explicitly takes between studies heterogeneity into account, has become the standard method in practice, though fixed effect methods are still being applied as well. For the simple case where meta-analysis concerns a single, univariate effect measure[2] the statistical methods are well established now. However, it is not rare that several outcome measures are presented in the individual studies included in a meta-analysis. In that case analyzing each outcome measure separately in a univariate manner is often sub-optimal, and analyzing all outcome measures jointly using multivariate methods is indicated. For many situations with multivariate outcome, appropriate meta-analytic methods are still lacking or underdeveloped. Van Houwelingen, Zwinderman and Stijnen[3] were the first to consider multivariate random effects meta-analysis. They introduced a bivariate linear random effects model for joint analysis of one estimated outcome measure per treatment group. Berkey[4] introduced the linear mixed model as a general random effects regression method for meta-analysis of multiple outcomes.

This thesis aims to be a contribution to the field of multivariate meta-analysis research. We considered four specific situations with multivariate outcome: a) the relationship between treatment effect and baseline risk (chapter 2), b) multivariate endpoints in a clinical trial (chapter 4), c) ROC curve data (chapter 5) and d) survival curve data (chapter 6). The statistical models that we used turned out to be very special cases of the general linear mixed model. Therefore we discussed in chapter 3 the general linear mixed model as a natural and convenient framework for meta-analysis. In the last 10 years programs for fitting linear mixed models have been implemented in many statistical packages. We discovered that these programs, provided that certain options are built in, could also be used for fitting meta-analysis models that are special cases of a general linear mixed model. An important contribution of chapter 3 to the practice of meta-analysis is that it thoroughly points out how many existing meta-analysis methods can be carried out using Proc Mixed of SAS, one of the most important statistical packages. Thus far ad hoc programs had to be used.

Using a linear mixed model for meta-analysis has its limitations. First, the within study likelihood of the outcomes is approximated by a normal likelihood. In particular if for a dichotomous outcome variable the number of events is close to 0 or 100 percent, this approximation might be bad, possibly introducing bias. To address this, a generalised linear mixed model might be used, in which the exact binomial likelihood is employed. However programs for the generalised linear mixed model are much more scarce and models are often difficult to fit in these programs. In the chapter on the relation of treatment effect and baseline risk we used the approximate normal as well as the exact binomial likelihood. The latter was carried out using the Bayesian analysis program BUGS. The results were quite similar. In chapter 6 on meta-analysis of ROC curve data we were able to fit the bivariate model with the exact binomial likelihood in Proc NLMIXED of SAS. In this case the difference between the results of the approximate and exact approach were not negligible. More research to this has to be done. At this moment we would advise to use the exact likelihood method whenever it is feasible.

A second limitation of the linear mixed model is that it is likelihood based, and thus the inference is approximate. This might be worrying when the number of studies included in the meta-analysis is small. More research has to be done as to whether this is a serious limitation. If the number of studies is small it may be better to fit the models using a Bayesian approach, since the Bayesian method is not asymptotic, but this might bring its own problems.

A third limitation of the general linear mixed model is that the random effects are assumed to have a multivariate normal distribution. How robust is the inference against violations of these assumptions? For instance, our bivariate approach in chapter 2 to investigating the relation between treatment effect baseline risk assumes that the underlying true baseline risks follow a normal distribution. This assumption has been criticised in the literature[5]. Thompson et al.[6], in their Bayesian approach, tried to avoid a distributional assumption for the underlying true baseline risks by putting independent flat priors on the true baseline risks. We criticized this method, but a simulation study comparing the relative merits of the two approaches is still lacking. Our approach is in the spirit of the 'structural' approach to measurement error[7]. As future research, a solution might be sought in the spirit of the 'functional' approach[7], the big advantage being that no distributional assumption for the true baseline risks is needed. Another direction of future research might be to try to estimate the underlying distribution of the baseline risks in a more non-parametric manner. The results of this research could also be useful for relaxing the assumptions in the bivariate model for ROC meta-analysis. A general remark is that in the Bayesian approach, for instance using WinBUGS, it relatively easy to relax the

assumption of a normal distribution of the random effects, by specifying other types of distributions such as *t*-distributions.

Another limitation of the linear mixed model is that it is not always sufficiently tailored to the nature of the problem. In chapter 5 we applied the linear mixed model to survival curve data, where clinical trials each present several survival percentages during a follow-up period. The advantage of using the linear mixed model is that it can be used for very unbalanced data sets and is easy to use in practice. The disadvantage however is that it does not completely satisfactorily take the nature of the data into account. The fitted curves are not necessarily survival curves. The model does not force the curves to be monotonically decreasing, and curves might start at values smaller than 100%. In future research, more elegant models should be sought that are better tailored to the nature of the data.

The methods for ROC curve meta-analysis presented in chapter 6 covered only the simplest situation, where one has one pair of sensitivity and specificity per study. Research on how to tackle more complicated situations, such as more points on the ROC curve per study or comparisons of different diagnostic tests with paired or unpaired data, has still to be done.

Of course, an unavoidable limitation of our methods is that they are just as susceptible as any meta-analysis to dangers that threat the validity of the conclusions, such as publication bias. This bias can arise when the meta-analysis does not contain all studies that fulfilled the in- and exclusion criteria, i.e. some studies are missing because they were not published or not traced. It can easily happen that the missing studies are selectively missing, for instance due to studies with a non-significant result having less chance to be published[8,9]. In this situation, a meta-analysis of the published trials could identify a spurious beneficial treatment effect, which cannot be prevented by our proposed statistical methods. Also, as in any meta-analysis, the complicated methods presented in this thesis do not help if the individual are biased or flawed. Still the law of 'garbage in garbage out' applies.

References

1. Stangl D, Berry D. *Meta-analysis in medicine and health policy.* Marcel Dekker, Inc.: New York, 2000.
2. Normand S-L. Meta-analysis: formulating, evaluating, combining and reporting. *Statistics in Medicine* 1999; **18**:321-359.
3. Van Houwelingen H, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; **12**:2272-2284.
4. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537-2550.
5. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal* 1996; **313**(7059):735-738.
6. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**(23):2741-2758.
7. Carroll RJ, Ruppert D, Stefanski LA. *Measurement error in nonlinear models*. Chapman & Hall: London, 1995.
8. Smith M. Publication bias and meta-analysis. *Evaluation Education* 1980; **4**:22-24.
9. Egger M, Smith G, Altman D. *Systematic reviews in health care. Meta-analysis in context.* BMJ Publishing Group: London, 2001.

Summary

Summary

Meta-analysis may be broadly defined as the quantitative review and synthesis of the results of related but independent studies. For the simple case where meta-analysis concerns only one outcome measure in each study, the statistical methods are well established now. However, in many practical situations there are several outcome measures presented in the individual studies included in a meta-analysis. In that case analyzing each outcome measure separately in a univariate manner is often sub-optimal, and analyzing all outcome measures jointly using multivariate methods is indicated. For many situations with multivariate outcome, appropriate meta-analytic methods are still lacking or underdeveloped. The work presented in this thesis aims at the development of statistical methods that are suited to analyse meta-analytic data with a multivariate nature in a right and optimal way. With our proposed methods one can answer more comprehensive research questions than with the standard univariate methods that are usually used in practice. Our explicit aim is that our suggested statistical methods are relatively easy to use for most researchers.

Chapter 1 is a general introduction to the topic of this thesis. In this chapter some basic terms from the field of meta-analysis are explained and an outline of the thesis is given.

The first situation of a meta-analysis with multivariate outcome in this thesis **(Chapter 2)** is the analysis of the relationship between treatment effect and baseline risk. A straightforward way of assessing this relationship is to compute the ordinary weighted least squares (WLS) regression line of the treatment effects estimated from the different trials on the estimated so called baseline risks observed in the control groups. This conventional method has potential pitfalls and has been seriously criticised. We propose another method based on a bivariate meta-analysis. Although we did most of the analyses using the BUGS implementation of Markov Chain Monte Carlo (MCMC) numerical integration techniques, we also show for one of the examples how it can be carried out with a general linear mixed model in SAS Proc Mixed. The advantage of using BUGS is that an exact measurement error model can be specified. On the other hand, in practice it is easier to use the procedure Proc Mixed of SAS.

In **Chapter 3** we discuss the general linear mixed model as a natural and convenient framework for meta-analysis. It is thoroughly pointed out how many existing meta-analysis methods can be carried out using Proc Mixed of SAS, one of the most important statistical packages. Thus far ad hoc programs had to be used. We discuss several methods to analyse univariate as well as bivariate outcomes in meta-analysis

and meta-regression analysis. Several extensions of the models are presented, like exact likelihood, non-normal mixtures and multiple endpoints. All methods are illustrated by a clinical meta-analysis example for which the complete syntax needed for the software program SAS is given.

In **Chapter 4** we discuss a meta-analysis of the effect of surgery (endarterectomy) compared to conservative treatment on the short and long term risk of stroke in patients with increased risk of stroke. Three summary measures per trial are available, which we jointly meta-analyse with a general linear mixed model. As far as we know this is the first published example of a multivariate random effects meta-analysis combining more than two outcomes simultaneously. We demonstrate the advantages of the multivariate analysis upon the univariate analyses where only one outcome measure at a time is measured. The multivariate approach reveals the relations between the different outcomes and gives simple expressions for estimation of derived treatment effect parameters such as the cumulative survival probability ratio as a function of follow-up duration. Besides, the results of the multivariate approach enable us also to estimate the relation of the different treatment effect parameters and the underlying risk. We fit the trivariate model in the standard general linear mixed model program of SAS using approximate likelihood. In a few special cases an exact likelihood approach is possible as well. For the multivariate case we fit the model using Bayesian methods to specify a binomial distribution for the number of post-operative events and a Poisson distribution for the events on long term in both treatment groups. The results of the approximate and exact likelihood approach are very similar.

Another application of multivariate meta-analysis in this thesis **(Chapter 5)** is the meta-analysis of ROC curve data. We consider the situation where per study on pair of estimated sensitivity and specificity is available. Meta-analysis of ROC-curve data is usually done with fixed effects models. Despite some random effects models have been published to execute a meta-analysis of ROC-curve data, these models are not often used in practice. Therefore we propose a more straightforward modelling technique for multivariate random effects meta-analysis of ROC-curve data, which can be fitted with standard software. The sensitivities and specificities of the diagnostic studies were analysed simultaneously using a two-dimensional random effects model. We show that different choices could be made to characterise the estimated bivariate normal distribution by a regression line or a so-called summary ROC curve. Under an extra assumption the model also provides individual study specific ROC curves. When a random intercept model is used to get individual study specific ROC curves, all study specific curves are parallel around the summary ROC curve. We have shown that it is also possible to fit a random slope next to a random

intercept, even when there is only one point per study. With the random intercept and slope model, the study specific ROC curves are not necessary parallel to the summary ROC curve anymore.

The general linear mixed model is also suited for meta-analysis of survival curve data **(Chapter 6)**, where clinical trials each present several survival percentages and their standard errors during a follow-up period. In practice the follow-up times and the number of follow-up times are different among studies. To tackle this problem, investigators often reduce the survival curve to one or some fixed points in time, e.g. the five-years survival rate. Then the data can be analysed with the standard univariate random (or fixed) effects model for each of the chosen time points separately. However, doing separate analyses for each point in time and thus carrying out many meta-analyses, is inefficient and could lead to inappropriate conclusions. Better methods have been proposed, but all of them are fixed effects methods. One of the most recommended methods is the one of Dear. Dear proposed a general linear model with survival estimate as dependent variable and follow-up time, treatment and study as categorical covariates. The parameters are estimated by generalised weighted least squares (GLS). In this thesis we generalise the GLS-method of Dear towards a multivariate random effects model, which could be applied on data with an arbitrary number of survival estimates and spacing of times between them per curve, possibly different between studies. This enables the meta-analyst to analyse all available data as provided in the publications, without need to inter- or extrapolate to fixed times. The method fits in principle in the framework of the general linear mixed model. However, it has to be adapted in this case, because the correlations between the different survival estimates of the same curve have to be estimated as well.

Finally, in **Chapter 7**, the main findings of this thesis are considered. In addition we discuss some limitations and make recommendations for future research.

Samenvatting

Samenvatting

Meta-analyse kan worden gedefinieerd als een kwantitatieve samenvoeging van de resultaten van gerelateerde maar onderling onafhankelijke onderzoeken. Als een meta-analyse gericht is op slechts één uitkomstmaat, dan is tegenwoordig duidelijk welke statistische methoden men moet gebruiken. Er zijn echter veel praktische situaties waar in de individuele studies meerdere uitkomstmaten worden gepresenteerd. In dat geval is het univariaat uitvoeren van een meta-analyse voor elke uitkomstmaat apart vaak niet optimaal. Het tegelijkertijd analyseren van alle uitkomstmaten met multivariate methoden is dan raadzaam.
Voor veel situaties waarin een meta-analyse een multivariate uitkomst heeft, ontbreken nog geschikte meta-analytische methoden of zijn deze nog onderontwikkeld. Dit proefschrift heeft als doel statistische methoden te ontwikkelen die geschikt zijn om meta-analytische data met een multivariaat karakter op een optimale manier te analyseren. Met de door ons voorgestelde methoden kan men veelomvattendere onderzoeksvragen beantwoorden dan met de standaard univariate methoden die normaal worden gebruikt. Ons doel is dat de door ons voorgestelde methoden relatief gemakkelijk te gebruiken zijn voor de meeste onderzoekers.

Hoofdstuk 1 geeft een inleiding over het onderwerp van dit proefschrift. In dit hoofdstuk worden bepaalde basistermen uit de meta-analyse uitgelegd en wordt een overzicht gegeven van het proefschrift.
De eerste toepassing van een meta-analyse met een multivariate uitkomst in dit proefschrift **(Hoofdstuk 2)** is de analyse van de relatie tussen het behandelingseffect en het onderliggende risico van de patiënten in het onderzoek. Een eenvoudige manier om deze relatie te berekenen, is door het berekenen van een gewogen kleinste kwadraten regressielijn, waarbij de behandelingseffecten in de verschillende studies worden geschat op basis van de geschatte 'onderliggende risico's'. Deze onderliggende risico´s worden berekend in de controle groepen van die studies. Deze conventionele methode heeft verschillende manco's en er is dan ook veel kritiek op deze methode is geleverd.
Wij stellen een andere methode voor, die gebaseerd is op een bivariate meta-analyse. De meeste analyses zijn met behulp van de BUGS implementatie van Markov Chain Monte Carlo (MCMC) numerieke integratie technieken gedaan. Voor één van de voorbeelden is daarnaast geïllustreerd hoe de methode kan worden uitgevoerd met een lineair gemengd model in SAS Proc Mixed. Het voordeel van BUGS is dat het exacte meetfouten model kan worden gespecificeerd. De procedure Proc Mixed van SAS is echter eenvoudiger in het gebruik.

In **Hoofdstuk 3** bespreken we het lineaire gemengde model als een eenvoudig en geschikt raamwerk voor meta-analyse. Er wordt geïllustreerd dat veel bestaande meta-analyse methoden kunnen worden uitgevoerd met Proc Mixed van SAS, één van de belangrijkste statistische software pakketten. Tot nu toe moesten daarvoor ad hoc programma's worden gebruikt.
Er komen verschillende methoden aan bod om univariate en bivariate uitkomsten in meta-analyse en meta-regressie analyse te analyseren. Verscheidene uitbreidingen van de modellen worden gepresenteerd, zoals exacte likelihood, gemengde verdelingen en meerdere eindpunten. Alle methoden worden geïllustreerd aan de hand van een meta-analyse voorbeeld ontleend aan de medische literatuur. Alle benodigde syntax voor het software programma SAS wordt volledig weergegeven.
In **Hoofdstuk 4** bespreken we een meta-analyse waar het effect van een operatie (carotis endarterectomie) wordt vergeleken met een medicinale behandeling. Gekeken wordt naar het korte en lange termijn effect van beide behandelingen op het krijgen van een beroerte bij patiënten die daar een verhoogd risico op hebben. Er zijn drie uitkomstmaten per studie beschikbaar, die we tegelijkertijd analyseren met een algemeen lineair gemengd model. Voor zover ons bekend, is dit het eerste gepubliceerde voorbeeld van een multivariate random effecten meta-analyse waarin meer dan twee uitkomsten samen worden geanalyseerd. We laten de voordelen zien van een multivariate analyse ten opzichte van univariate analyses waarin elke uitkomst apart wordt geanalyseerd. De multivariate benadering laat de relaties tussen de verschillende uitkomsten zien en geeft eenvoudige functies voor het schatten van afgeleide behandelingseffecten zoals de cumulatieve overlevingskans ratio als een functie van de tijd dat de patiënten na de behandeling in het onderzoek zijn gevolgd. Bovendien maken de resultaten van een multivariate benadering het mogelijk om ook de relatie te schatten tussen de verschillende behandelingseffectparameters en het onderliggende risico. We hebben bij het 'fitten' van het trivariate model in het standaard lineaire gemengd model programma van SAS gebruik gemaakt van benaderende ('approximate') likelihood. In een paar speciale gevallen is ook het gebruik van de exacte likelihood mogelijk. Voor de multivariate situatie hebben we het model gefit met Bayesiaanse methoden om de binomiale verdeling te specificeren van het aantal beroertes na de operatie en een Poisson verdeling te specificeren voor het aantal beroertes op de lange termijn in beide behandelingsgroepen. De resultaten van het gebruik van de benaderende ('approximate') en de exacte likelihood lijken erg op elkaar.
Een andere toepassing van multivariate meta-analyse in dit proefschrift **(Hoofdstuk 5)** is de meta-analyse van ROC curve data. We beschouwen de situatie waar per studie één paar geschatte sensitiviteit en specificiteit beschikbaar is. Meta-analyse van

ROC curve data wordt gewoonlijk gedaan met vaste effecten modellen. Ondanks dat er enkele random effect modellen zijn gepubliceerd om een meta-analyse van ROC curve data uit te voeren, worden deze modellen in de praktijk niet vaak toegepast, omdat deze nogal complex zijn. Wij stellen een eenvoudiger model voor multivariate random effecten meta-analyse van ROC-curve data voor dat kan worden uitgevoerd met standaard software. De sensitiviteiten en de specificiteiten van de diagnostische studies worden samen geanalyseerd met een tweedimensionaal random effecten model. We laten zien dat er verschillende keuzes kunnen worden gemaakt om de geschatte bivariate normale verdeling te karakteriseren door een regressielijn of een zogenaamde samengevatte ROC curve. Onder een extra aanname biedt het model bovendien individuele studie-specifieke ROC curves. Als een random intercept model wordt gebruikt om individuele studie-specifieke ROC curves te verkrijgen, dan zijn alle studie-specifieke curves parallel aan de samengevatte ROC curve. We laten zien dat het ook mogelijk is om een model te fitten met een random intercept èn een random helling, zelfs als er maar één punt op de ROC curve per studie gegeven is. Met een model met een random intercept en random helling zijn de studie-specifieke ROC curves niet meer noodzakelijkerwijs parallel aan de samengevatte ROC curve.

Het algemene lineaire gemengde model is ook geschikt voor meta-analyse van overlevingsduur data **(Hoofdstuk 6)**, waarbij elke klinische trial meerdere survival percentages presenteert met de bijbehorende standaardfouten gedurende een bepaalde periode nadat de behandeling is gestart. In de praktijk zijn de tijden waarop de gegevens zijn gemeten en het aantal metingen per studie zeer verschillend. Om dit probleem te omzeilen, reduceren onderzoekers vaak de overlevingscurve tot één of meerdere vaste tijdstippen, bijvoorbeeld de overleving na vijf jaar. In dat geval kunnen de data voor elk van de gekozen tijdstippen apart worden geanalyseerd met het standaard univariate random (of vaste) effecten model. Het uitvoeren van aparte analyses voor elk tijdstip en dus het uitvoeren van veel meta-analyses, is echter inefficiënt en kan leiden tot foute conclusies. Er zijn betere methoden voorgesteld, maar dat zijn allemaal vaste effecten modellen. Eén van de meest aanbevolen (vaste effecten) methoden is het model van Dear. Dear stelde een algemeen lineair model voor met de schattingen van de overlevingspercentages als afhankelijke variabele en met de follow-up tijd, behandeling en studie als categorische covariaten. De parameters worden geschat met GLS (generalised weighted least squares). In dit proefschrift hebben we de GLS-methode van Dear uitgebreid tot een multivariaat random effecten model, dat kan worden toegepast op data met een willekeurig aantal overlevingsschattingen en waarbij willekeurige tijdsintervallen tussen de schattingen bestaan, mogelijk verschillend tussen de studies. Dit maakt het voor degene die de meta-analyse doet mogelijk om alle beschikbare data te analyseren zoals ze worden

weergegeven in de publicaties, zonder de noodzaak van inter- of intrapolatie naar vaste tijdstippen. De methode past in het raamwerk van het algemene lineaire gemengde model. Het moet echter wel voor deze specifieke situatie worden aangepast, omdat de correlaties tussen de verschillende overlevingsschattingen ook moeten worden geschat.

Tenslotte worden in **Hoofdstuk 7** de belangrijkste bevindingen van dit proefschrift beschouwd. Bovendien bespreken we enkele beperkingen van de voorgestelde modellen en doen we aanbevelingen voor verder onderzoek.

Dankwoord

About the author

List of publications

Dankwoord

Het meest gelezen hoofdstuk uit het proefschrift is aangebroken. Zoals de meeste mensen om mij heen zullen hebben meegekregen, was dit boekje er nooit geweest zonder twee personen. Allereerst ben ik mijn promotor Theo Stijnen veel dank verschuldigd voor het feit dat ik altijd mocht binnen lopen met vragen en voor al zijn ideeën en oplossingen. Beste Theo, bedankt ook voor al je begrip en geduld in moeilijkere tijden, een prettiger mens kon ik me niet wensen als promotor. Ik hoop ook in de toekomst nog veel van je te mogen leren. De andere persoon is Geeske Hofstra. Na onze studie Econometrie koos zij de voor mij geschikte baan als biostatisticus in Rotterdam, terwijl ik voor een andere baan had gekozen. Na twee jaar ging Geeske naar Den Haag en kon ik haar baan overnemen. Beste Geeske, zonder jou had ik deze leuke baan en deze leuke collega´s niet gehad. Ik ben er trots op dat jij tijdens de promotie mijn paranimf wilt zijn.

Henk Schmidt en Bert Hofman wil ik ervoor bedanken dat ik zowel bij Psychologie als bij Epidemiologie & Biostatistiek mag werken. In beide omgevingen voel ik me als een vis in het water. Ik ervaar het als een enorme luxe om twee zulke leuke werkplekken te hebben.

De co-auteurs en tevens leden van de kleine commissie Hans van Houwelingen en Myriam Hunink wil ik graag bedanken voor hun suggesties voor verbeteringen van de artikelen.

Een promotietraject verloopt een stuk beter in een aangenaam werkklimaat. Mijn kamergenoten Maria en Bettina bij Biostatistiek en Eveline bij Psychologie, bedankt voor de geweldige sfeer op onze kamer, ik verheug me er altijd op jullie weer te zien. Maria, ik ben erg blij dat jij mij als paranimf tijdens de promotie terzijde wilt staan. Bettina, bedankt dat jij bereid bent om als derde paranimf op te treden, ook al is dat een rol achter de schermen. Eveline, bedankt voor alle thee, paaseitjes en gezelligheid. Ik hoop nog heel lang met jullie een kamer te mogen delen.

Ron Olij ben ik zeer erkentelijk voor zijn prachtige ontwerp voor de omslag en voor zijn tips voor de layout. Dankzij Eveline is ook de binnenkant zeer fraai geworden.

Nano, dank voor alle keren dat je mij hielp als mijn computer niet deed wat ik wilde.

Alle collega's van Biostatistiek (met name Wim en Paul), Epidemiologie (met name Jacqueline, Majanka en Joke) en natuurlijk van Psychologie (teveel om op te noemen!) ben ik dankbaar voor alle gezelligheid en meelevendheid.

Tegen al mijn vriendinnen, vrienden en (schoon-)familie kan ik oprecht zeggen dat ik niet zonder jullie kan. Bedankt voor jullie betrokkenheid, enthousiasme, openheid en voor jullie steun. Lieve José, dank voor alle herinneringen. Ik mis je.

Lieve Wim, zonder jou was dit boekje vast veel eerder af geweest ;-)). Lieve Wouter en Suzanne, ook jullie maken me elke dag duidelijk dat het leven uit zoveel meer bestaat dan werk alleen. Laten we heel oud en gelukkig worden met z'n allen.

About the author

Lidia Arends was born April 20th, 1970 in Eelde. In 1988 she obtained the diploma VWO-beta at St. Maartens-college in Haren (Gr.). She graduated in 1994 at the University of Groningen in Econometrics (cum laude) and in Psychology (cum laude).
That same year she joined the Netherlands Institute of Mental Health (Trimbos Institute) in Utrecht as a research scientist in the field of cost-effectiveness of the mental health care under guidance of Prof.dr. M.C.H. Donker. In 1996 she obtained a Master of Science in Health Services Research at the Netherlands Institute of Health Sciences (NIHES) in Rotterdam.
Since 1997 she started the work described in this thesis under guidance of Prof.dr. Th. Stijnen at the Department of Epidemiology & Biostatistics (head: Prof.dr. A. Hofman) of the Erasmus Medical Center in Rotterdam, where she became biostatistician. In 2001 she obtained a Master of Science in Epidemiology at the NIHES. She is registrated as Biostatistician VVS since March 2002.
Since February 2002 she also holds a position as assistant professor at the Department of Psychology, Erasmus University Rotterdam (head: Prof.dr. H.G. Schmidt), where she became involved in the development of a new academic psychology curriculum. She is engaged in developing and conducting the statistics courses.
Together with Wim Brunenberg she has two children, Wouter and Suzanne.

List of publications

Arends LR. *Vergelijking van het Peabody & Goldbergmodel met het AB5C-model aan de hand van chiasmen*. Heymans Bulletin, HB-94-1138-SW: Rijksuniversiteit Groningen, 1994.

Hofstee WKB, Arends LR. The heuristic potential of the Abridged Big-Five Dimensional Circumplex (AB5C) model: Explaining the chiasmic illusion. *Psychologica Belgica* 1994; **34**(4):195-206.

Arends LR. *Van koude gatenkaas naar warme schatters. Attitudeschalen betreffende kwaliteit van dienstverlening ondanks ontbrekende waarnemingen in hun onderlinge relaties bestudeerd*. ITB-RAP-94-758, PTT Research/ Instituut voor Toegepast Bedrijfsonderzoek: Groningen, 1994.

Stoop B, Arends L, van Leeuwen R. Development of an instrument for determining organisational service quality. *Research Review* 1995; **5**(4):4-27.

Van Roijen L, Arends LR. *Kosten van psychische ziekten*. iMTA/NcGv: Rotterdam, 1996.

Donker M, Arends L, Gageldonk A van, Roijen L van, Smit F, Bijl R. *Prioriteiten voor kosten-effectiviteitsonderzoek in de GGZ*. NcGv-reeks 96-15: Utrecht, 1996.

Arends L, van Gageldonk A. *Het effect van interenties bij stemmingsstoornissen. Overzicht van meta-analyses en reviews*. Trimbos-instituut: Utrecht, 1997.

Van Dam QD, Van 't Spijker A, Arends LR, Hakkaart-van Roijen L, Donker MCH, Trijsburg RW. Doelmatigheid van psychotherapie: een haalbaarheidsonderzoek. *Tijdschrift voor Psychotherapie* 1998; **24**(1):5-22.

Stijnen T, Arends LR. Dwalingen in de methodologie. XVI. Wat te doen met missing values? *Nederlands Tijdschrift voor Geneeskunde* 1999; **143**(40):1996-2000.

Meijer RJ, Kerstjens HA, Arends LR, Kauffman HF, Koeter GH, Postma DS. Effects of inhaled fluticasone and oral prednisolone on clinical and inflammatory parameters in patients with asthma. *Thorax* 1999; **54**(10):894-899.

Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine* 2000; **19**(24):3497-3518.

Kasteleijn-Nolst Trenite DGA, Rentmeester ThW, Scholtes FBJ, Gilissen KGPM, Arends LR, Schlosser A. Peri-Marketing Surveillance of lamotrigine in the Netherlands - Doctors' and patients' viewpoints. *Pharmacy World & Science* 2001; **23**(1):1-5.

Visser MJ, Postma DS, Arends LR, de Vries TW, Duiverman EJ, Brand PL. One-year treatment with different dosing schedules of fluticasone propionate in childhood asthma. Effects on hyperresponsiveness, lung function, and height. *American Journal of Respiratory and Critical Care Medicine* 2001; **164**(11):2073-2077.

Van Houwelingen H, Arends LR, Stijnen T. Tutorial in Biostatistics. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**: 589-624.

Meijer RJ, Postma DS, Kauffman HF, Arends LR, Koeter GH, Kerstjens HA. Accuracy of eosinophils and eosinophil cationic protein to predict steroid improvement in asthma. *Clinical and Experimental Allergy* 2002; **32**:1096-1103.

Visser MJ, Postma DS, Brand PLP, Arends LR, Duiverman EJ, Kauffman HF. Influence of different dosage schedules of inhaled fluticasone propionate on peripheral blood cytokine concentrations in childhood asthma. *Clinical and Experimental Allergy* 2002. **32**:1497-1503.

Carpay JA, Linssen WHJP, Koehler PJJ, Arends LR, Tiedink HGM. Efficacy of sumatriptan nasal spray in recurrent migrainous headaches (IHS 1.7): an open, prospective study. *Headache* 2003; **43**(4):395-399.

van Buuren HR, Rasch MC, Batenburg PL, Bolwerk CJM, Nicolai JJ, van der Werf SDJ, Scherpenisse J, Arends LR, van Hattum J, Rauws EAJ, Schalm SW. Endoscopic sclerotherapy compared with no specific treatment for the primary prevention of bleeding from esophageal varices. A randomized controlled multicentre trial [ISRCTN03215899]. *BMC Gastroenterology* 2003; **3**:22-32.

Arends LR, Vokó Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine* 2003; **22**(8):1335-1353.

Streppel MT, Arends LR, Grobbee DE, van 't Veer P, Geleijnse JM. Effect of fiber supplementation on blood pressure. *Circulation* 2004; **109**(7):E136-E136 P275.

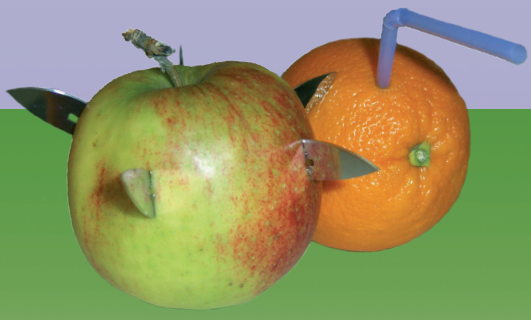
Visser MJ, van der Veer E, Postma DS, Arends LR, de Vries TW, Brand PL, Duiverman EJ. Side-effects of fluticasone in asthmatic children: no effects after dose reduction. *European Respiratory Journal* 2004; **24**(3):420-425.

Twilt M, Mobers SM, Arends LR, ten Cate R, van Suijlekom-Smit L. Temporomandibular involvement in juvenile idiopathic arthritis. *Journal of Rheumatology* 2004; **31**(7):1418-1422.

Streppel MT, Arends LR, van't Veer P, Grobbee DE, Geleijnse JM. Dietary fiber and blood pressure - A meta-analysis of randomized placebo-controlled trials. *Archives of Internal Medicine* 2005; **165**(2): 150-156.

Noordzij M, Uiterwaal CSPM, Arends LR, Kok FJ, Grobbee DE, Geleijnse JM. Blood pressure response to chronic intake of coffee and caffeine: a meta-analysis of randomized controlled trials. *Journal of Hypertension* 2005; **23**(5):921-928.

van Mierlo LAJ, Arends LR, Streppel MT, Zeegers MPA, Kok FJ, Grobbee DE, Geleijnse JM. Blood pressure response to calcium supplementation: a meta-analysis of randomized controlled trials. *Journal of Human Hypertension* 2006; **5**:1-10.



ISBN 90 90 20786 4