META-ANALYSIS

New Developments and Applications in Medical and Social Sciences

META-ANALYSIS

New Developments and Applications in Medical and Social Sciences

Edited by

Ralf Schulze, Heinz Holling, & Dankmar Böhning

Hogrefe & Huber

Library of Congress Cataloging-in-Publication Data

is now available via the Library of Congress Marc Database under the

LC Control Number: 2003104139

National Library of Canada Cataloguing-in-Publication

Meta-analysis: new developments and applications in medical and social sciences / Ralf Schulze, Heinz Holling, Dankmar Böhning (editors).

Includes bibliographical references and index.
ISBN 0-88937-266-7

1. Meta-analysis. 2. Medicine--Research--Evaluation. 3. Clinical trials--Statistical methods. 4. Social sciences--Research--Evaluation. 5. Social sciences--Statistical methods. I. Schulze, Ralf II. Holling, Heinz III. Böhning, Dankmar

R853.S7M48 2003 610'.7'27 C2003-902748-1

PUBLISHING OFFICES
USA: Hogrefe & Huber Publishers, 44 Brattle Street, 4th Floor,
Cambridge, MA 02138
Phone (866) 823-4726, Fax (617) 354-6875, E-mail info@hhpub.com
Europe: Hogrefe & Huber Publishers, Rohnsweg 25, D-37085 Göttingen, Germany,
Phone +49 551 49609-0, Fax +49 551 49609-88, E-mail hh@hhpub.com

SALES & DISTRIBUTION
USA: Hogrefe & Huber Publishers, Customer Services Department,
30 Amberwood Parkway, Ashland, OH 44805,
Phone (800) 228-3749, Fax (419) 281-6883, E-mail custserv@hhpub.com
Europe: Hogrefe & Huber Publishers, Rohnsweg 25, D-37085 Göttingen, Germany,
Phone +49 551 49609-0, Fax +49 551 49609-88, E-mail hh@hhpub.com

OTHER OFFICES
Canada: Hogrefe & Huber Publishers, 12 Bruce Park Avenue, Toronto, Ontario M4P 2S3
Switzerland: Hogrefe & Huber Publishers, Länggass-Strasse 76, CH-3000 Bern 9

Hogrefe & Huber Publishers
Incorporated and registered in the State of Washington, USA, and in Göttingen, Lower Saxony, Germany

Printed and bound in Germany
ISBN 0-88937-266-7

Preface

Meta-analysis as a systematic method to integrate empirical findings has become a widely adopted technique in various scientific fields. Among the major areas of application of the method are medicine and the social sciences. New statistical developments and methodological advances often happen unrecognized in different substantive fields, or are assimilated with considerable delay. The present volume is intended to bring scholars from medical and social sciences together to present their theoretical advances as well as new applications of the method.

The book is divided in two parts. The first part consists of a collection of chapters that address various important theoretical issues. These chapters focus on the evaluation and systematization of existing procedures that are used in practice, present new developments regarding statistical procedures, describe techniques for the detection of bias in meta-analysis, and provide detailed expositions of the methodological viewpoints on meta-analysis in pharmaceutical, medical as well social science research.

In Chapter 1, Hartung, Argaç, and Makambi present a series of homogeneity tests that are known within the framework of ANOVA but have not been widely adopted in applications of meta-analysis. They expound the underlying logic of the tests and evaluate their performance in a simulation study. Hartung et al. address the problem of testing the homogeneity assumption that is often made in practical applications of meta-analysis, and they show which tests perform best under several conditions.

Schulze, Holling, Großmann, Jütting, and Brocke present a comparison of two meta-analytical approaches for the analysis of correlation coefficients in Chapter 2. It is shown that parallel statistical developments in different subdisciplines of psychology have lead to diverse procedural details in approaches often used in practice. These details can in turn lead to differences in results on the basis of the same database. This is demonstrated in a Monte-Carlo study of different homogeneous situations for which the procedures of the approaches – and fixed effects models in general – are supposed to be appropriate.

Random and fixed effects models in meta-analysis play an important role for the data analytic strategy and the interpretation of results. In recent years, the random effects model has been favored over the fixed effects model for theoretical reasons but only few procedures have been proposed for the estimation of the heterogeneity variance. This variance is an important component in the random effects model. Malzahn presents a general principle for its estimation in several meta-analytical models in Chapter 3.

The choice between the random and fixed effects model of meta-analysis has been subject of several debates. Although the random effects model was focused in theoretical discussions of the topic, in practical applications of meta-analysis, especially in the social sciences, the fixed effects model still prevails (see Chapter 2). Several authors have argued that the choice between these models has to be based on theoretical reasons and the inference that is intended with a meta-analysis. Hartung and Knapp present the basics of both the random and fixed effects model as well as commonly used methods in these models in Chapter 4. They also show that there are *theoretical* deficiencies in these models and propose an alternative test procedure which is presented in detail from an analytical point of view. Furthermore, the results of a simulation study that evaluates the performance of this new test procedure is reported.

The issue of bias in meta-analysis poses considerable problems to the interpretation of meta-analytical results. Often, the so-called publication bias is of particular interest. In Chapter 5, Schwarzer, Antes, and Schumacher review several procedures – graphical methods as well as test procedures – for the detection of bias in meta-analysis. They also present the results of a simulation study to evaluate the performance of two statistical tests for the identification of bias.

Apart from statistical issues in a narrower sense like those addressed in the first five chapters, more general methodological discussions have reoccurred in the literature since the advent of meta-analysis. Such methodological issues are addressed in the following four chapters. The different perspectives of medical research and the social sciences are reflected in these chapters and it is shown how analogous problems are dealt with in these areas of research.

In Chapter 6, Sauerbrei and Blettner review and compare different methods for summarizing empirical results from observational studies, including narrative reviews, meta-analysis of literature, meta-analysis of patient data, and prospective meta-analysis. Focusing on applications to medical research problems, the utility of meta-analysis for the evaluation of medical treatments is critically assessed. In addition to a theoretical analysis of the different review methods, several examples from the medical literature are presented. These examples support their arguments for a sceptical view on the utility of meta-analyses that are based on summary reports from the literature.

Koch and Röhmel concentrate in Chapter 7 on the use of meta-analysis in the process of new drug applications, where the method has not played a major role to date. They point out obstacles for the acceptance of meta-analytical results in this area. An analysis of the evaluation process for outcomes from randomized clinical trials on the comparison of different drugs for the same indication is presented, and references to relevant guidelines are given. Also, problems as well as benefits in using meta-analysis are illustrated by giving concrete examples. The characteristics that influence the credibility of meta-analyses in this field of application are highlighted as well. Thereby, Koch and Röhmel provide a constructive account for the enhancement of meta-analytical design.

In the subsequent chapter, Matt presents a comprehensive treatment on the possibilities to draw generalized causal inferences based on the results of meta-analysis. Here, like in other chapters in this volume, it is acknowledged that methods of meta-analysis are comparable to quasi-experiments or observational studies in methods of primary research. Drawing on principles developed in the context of generalization in quasi-experimentation, he demonstrates how these principles can be fruitfully applied to methods of meta-analysis. In his detailed exposition Matt also refers to general principles of generalization and provides examples of their successful application in practice. The presentation in Chapter 8 by Matt shows how questions of generalization are treated in the social sciences, and this view stands – at least partly – in contrast to treatments from the perspective of medical research (see e.g., Chapter 6 by Sauerbrei and Blettner).

The last chapter of the first part addresses the utility of tests of moderator hypotheses in meta-analysis. In Chapter 9 by Czienskowski, an example from social cognition research on the so-called self-reference effect is given to illustrate the application of moderator-analysis. Potential conclusions on the basis of the results are discussed, and it is shown how and why moderator analyses can and should be supplemented by follow-up experiments.

In the second part of the book applications of meta-analysis to different problems in medical, pharmaceutical and social science research are presented. A series of six chapters illustrates the breath of potential fields of application for meta-analytic methods.

An innovative field of application for meta-analysis is quality control in pharmaceutical production. In Chapter 10, Böhning and Dammann provide an overview and an example on how methods of meta-analysis can be applied in this new area of application. They extend an approach of mixture modeling of heterogeneity in meta-analysis and show its potential for an improvement of production processes in pharmaceutical industry.

In the following Chapter 11 by Greiner, Wegscheider, Böhning, and Dahms, an application of meta-analysis to explore and identify factors that influence the sensitivity and specificity of a medical test for the detection of trichinella antibodies is presented. They illustrate how adequate statistical methods of meta-analysis (e.g., mixed logistic regression) can contribute new knowledge that is of practical concern.

In Chapter 12, Dietz and Weist introduce a method based on finite mixed generalized linear models as a means for modeling heterogeneity in meta-analytic data. They present a detailed account of the model, methods for the estimation of parameters, and also give two examples of its application. The authors thereby demonstrate how advanced flexible methods of meta-analysis can provide useful results for the explanation of heterogeneity that go well beyond information gained from ordinary applications of meta-analysis.

Franklin also uses the generalized linear model in Chapter 13 to assess the impact of explanatory variables on the variability in a meta-analytical database. He examines, among other influential factors, the differences between treatment results in paediatric and adult clinical trials on Hodgkin's disease.

In a meta-analysis on the results of controlled clinical trials on antidepressants, Schöchlin, Klein, Abrahm-Rudolf, and Engel examine the potential moderating influence of design variables. They report results in Chapter 14 that stress the important role of design variables – especially the inclusion of placebo conditions – in this area of clinical applications.

One of the major research fields in social psychology, attitude research, is the subject of Chapter 15 by Schulze and Wittmann. The authors first provide an exposition of the two most often applied theories in this area. Additionally, moderator hypotheses concerning the relationships between the theory's components are substantiated that reflect standard assumptions of the theories as well as new hypotheses not previously tested in a meta-analytical framework. The results of a meta-analysis are also presented to assess overall effects as well as tests of pertinent moderator hypotheses in a random effects model.

Finally, Schlattmann, Malzahn, and Böhning present a new software package called META for the application of meta-analysis in Chapter 16. META enables the user to perform not only standard analysis to integrate research results but also includes procedures to apply the latest developments in mixture modeling of heterogeneity in meta-analysis as presented in this volume (see also Chapter 10).

The new developments and applications described in these chapters are contributions from different fields of research. Our hopes are that bringing together the contributions from these scholars in a single volume adds new knowledge to the different fields, counteracts fragmentation of statistical and substantial developments, and encourages potential users of the procedures to apply the latest methods of meta-analysis in their field of interest.

RALF SCHULZE
HEINZ HOLLING
DANKMAR BÖHNING

Contents

Part I

Theory

1
Homogeneity Tests in Meta-Analysis

Joachim Hartung
Doğan Argaç
Department of Statistics[†]
University of Dortmund

Kepher Makambi
Department of Mathematics and Statistics
Jomo Kenyatta University of Agriculture and Technology

Summary

For the homogeneity problem in meta-analysis, the performance of seven test statistics is compared under homogeneity and heterogeneity of the underlying population (study, group) variances. These are: the classical ANOVA *F* test, the Cochran test, the Welch test, the Brown-Forsythe test, the modified Brown-Forsythe test, the approximate ANOVA *F* test and as a proposal, an adjusted Welch test. At the whole, the Welch test proves to be the best one, but for small sample sizes and many groups, it becomes too liberal. In this case the adjusted Welch test is recommended to correct this anomaly. The other tests prove to have changing advantages dependent on the sizes of the parameters involved.

[†] Project "Meta-Analysis in Biometry and Epidemiology" (SFB 475) of the Deutsche Forschungsgemeinschaft (DFG).

1.1 INTRODUCTION

Meta-analysis of results from different experiments (groups, studies) is a common practice nowadays. In the framework of a one-way ANOVA model, serving generally as supporting edifice for meta-analysis, one may be interested in testing the homogeneity hypothesis. However, when the underlying population variances in different populations (studies, groups) are different, the ANOVA F-statistic attains significance levels which are very different from the nominal level (see for example, De Beuckelaer, 1996). In the rubric of the (generalized) Behrens-Fisher problem, a number of alternatives have been suggested.

Using simulation studies for various constellations of number of populations, sample sizes and within population error variances, we compare the actual attained sizes of the classical ANOVA F test, the Cochran test, the Welch test, the Brown-Forsythe test, the modified Brown-Forsythe test, the approximate ANOVA F test and, by adopting an idea of Böckenhoff and Hartung (1998), an adjusted Welch test, simultaneously.

1.2 MODEL AND TEST STATISTICS

Let y_{ij} be the observation on the jth subject of the ith population/study, $i = 1,\ldots,K$ and $j = 1,\ldots,n_i$

$$\begin{aligned} y_{ij} &= \mu_i + e_{ij} \\ &= \mu + a_i + e_{ij}\,; \quad i = 1,\ldots,K,\ j = 1,\ldots,n_i, \end{aligned}$$

where μ is the common mean for all the K populations, a_i is the effect of population i with $\sum_{i=1}^K a_i = 0$, and e_{ij}, $i = 1,\ldots,K, j = 1,\ldots,n_i$ are error terms which are assumed to be mutually independent and normally distributed with

$$E(e_{ij}) = 0, \quad \mathrm{Var}(e_{ij}) = \sigma_i^2\,;\ i = 1,\ldots,K,\ j = 1,\ldots,n_i$$

That is, $e_{ij} \sim \mathcal{N}(0,\sigma_i^2)$; $i = 1,\ldots,K,\ j = 1,\ldots,n_i$.

Interest is in testing the hypothesis $H_0 : \mu_1 = \cdots = \mu_K = \mu$. To test this hypothesis we will make use of the following test statistics:

a) The ANOVA *F* Test

S_{an}, given by

$$S_{an} = \frac{N-K}{K-1} \cdot \frac{\sum_{i=1}^K n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^K (n_i - 1)s_i^2}, \tag{1.1}$$

with $N = \sum_{i=1}^K n_i$, $\bar{y}_{i.} = \sum_{j=1}^K y_{ij}/n_i$, $\bar{y}_{..} = \sum_{i=1}^K n_i\bar{y}_{i.}/N$.

This test was originally meant to test for equality of population means under variance homogeneity and has an F distribution with $K-1$ and

$N - K$ degrees of freedom.

Test: Reject $H_0 : \mu_1 = \cdots = \mu_K$ at level α if $S_{an} > F_{K-1,N-K;1-\alpha}$.

The ANOVA test has the weakness of not being robust with respect to heterogeneity in the intra-population error variances (Brown & Forsythe, 1974).

b) The Welch Test

$$S_{we} = \frac{\sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2}{\left((K-1) + 2 \cdot \frac{K-2}{K+1} \cdot \sum_{i=1}^{K} \frac{1}{n_i - 1}(1-h_i)^2\right)}, \tag{1.2}$$

where $w_i = n_i / s_i^2, \quad h_i = w_i / \sum_{k=1}^{K} w_k$, was an extension of testing the equality of two means to more than two means (see Welch, 1951) in the presence of variance heterogeneity within populations.

Under H_0, the statistic S_{we} has an approximate F distribution with $K - 1$ and ν_g degrees of freedom, where

$$\nu_g = \frac{(K^2 - 1)/3}{\sum_{i=1}^{K} \frac{1}{n_i - 1}(1-h_i)^2}.$$

Test: Reject H_0 at level α if $S_{we} > F_{K-1,\nu_g;1-\alpha}$.

c) Cochran's Test

$$S_{ch} = \sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2, \tag{1.3}$$

was proposed by Cochran (1937) and then modified by Welch. We take it into our comparisons in order to get better comprehension and insight of the behavior of both statistics.

Under H_0, the Cochran statistic is distributed approximately as a χ^2-variable with $K - 1$ degrees of freedom.

Test: Reject H_0 at level α if $S_{ch} > \chi^2_{K-1;1-\alpha}$.

d) Brown-Forsythe (B-F) Test

This one is also known as the modified F test and is given by

$$S_{b-f} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}. \tag{1.4}$$

When H_0 is true, S_{b-f} is distributed approximately as an F variable with $K-1$ and ν degrees of freedom where

$$\nu = \frac{\left(\sum_{i=1}^{K}(1 - n_i/N)s_i^2\right)^2}{\sum_{i=1}^{K}(1 - n_i/N)^2 s_i^4/(n_i - 1)}. \tag{1.5}$$

Test: Reject H_0 at level α if $S_{b-f} > F_{K-1,\nu;1-\alpha}$.

Using a simulation study Brown and Forsythe (1974) demonstrated that their statistic is robust under inequality of variances. If the population variances are homogeneous, the B-F test is closer to ANOVA than Welch.

e) Mehrotra (Modified Brown-Forsythe) Test

$$S_{b-f(m)} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}, \tag{1.6}$$

was proposed by (Mehrotra, 1997) in an attempt to correct a "flaw" in the B-F test.

Under H_0, $S_{b-f(m)}$ is distributed approximately as an F variable with ν_1 and ν degrees of freedom where

$$\nu_1 = \frac{\left(\sum_{i=1}^{K}(1 - n_i/N)s_i^2\right)^2}{\sum_{i=1}^{K} s_i^4 + \left(\sum_{i=1}^{K} n_i s_i^2/N\right)^2 - 2 \cdot \sum_{i=1}^{K} n_i s_i^4/N} \tag{1.7}$$

and ν is given in Equation 1.5 above.

Test: Reject H_0 at level α if $S_{b-f(m)} > F_{\nu_1,\nu;1-\alpha}$.

The flaw mentioned above is in the estimation of the numerator degrees of freedom by $K-1$ instead of ν_1.

f) The Approximate ANOVA *F* Test

$$S_{aF} = \frac{N-K}{K-1} \cdot \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(n_i - 1)s_i^2}, \tag{1.8}$$

by Asiribo and Gurland (1990). This test gives an approximate solution to the problem of testing equality of means of normal populations in case of heteroscedasticity by making use of the classical ANOVA test.

Under H_0, the statistic S_{aF} is distributed approximately as an F-variable with ν_1 and ν_2 degrees of freedom where ν_1 is as given in Equation 1.7 above and

$$\nu_2 = \frac{\left(\sum_{i=1}^{K}(n_i - 1)s_i^2\right)^2}{\sum_{i=1}^{K}(n_i - 1)s_i^4}. \tag{1.9}$$

Test: Reject H_0 at level α if $S_{aF} > \hat{c} \cdot F_{\nu_1,\nu_2;1-\alpha}$, where

$$\hat{c} = \frac{N-K}{N(K-1)} \frac{\sum_{i=1}^{K}(N - n_i)s_i^2}{\sum_{i=1}^{K}(n_i - 1)s_i^2}. \tag{1.10}$$

We notice that the numerator degrees of freedom for S_{aF} and $S_{b-f(m)}$ are equal. Further, for $n_i = n, i = 1, \ldots, K$, that is, for balanced samples, the test statistic and the degrees of freedom for both the numerator and denominator of these two statistics are also equal. That is, for balanced designs

$$S_{aF} = S_{b-f(m)} = \frac{nK}{K-1} \cdot \frac{\sum_{i=1}^{K}(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K} s_i^2},$$

and

$$\nu = \nu_2 = (n-1) \cdot \frac{\left(\sum_{i=1}^{K} s_i^2\right)^2}{\sum_{i=1}^{K} s_i^4}.$$

g) The Adjusted Welch Test

The Welch Test uses weights $w_i = n_i/s_i^2$. We know that

$$E(w_i) = E\left(\frac{n_i}{s_i^2}\right) = c_i \cdot \frac{n_i}{\sigma_i^2},$$

where $c_i = (n_i - 1)/(n_i - 3)$, see Patel, Kapadia, and Owen (1976, pages 39-40). Therefore, an unbiased estimator of n_i/σ_i^2 is $n_i/c_i s_i^2$.

Now, let $\varphi_i = (n_i + \delta_1)/(n_i + \delta_2)$, where δ_1 and δ_2 are arbitrary real numbers; and then define the general weights by $w_i^* = n_i/\varphi_i s_i^2$. That is, for the Welch test, $w_i = w_i^*$ with $\varphi_i = 1$ ($\delta_1 = 0$, and $\delta_2 = 0$) and if we take the unbiased weights, $w_i = n_i/c_i s_i^2$, then $\varphi_i = c_i$, ($\delta_1 = -1$ and $\delta_2 = -3$).

For small samples in the groups, the Welch test becomes too liberal especially with increasing number of groups. Also, in our experience, using the unbiased weights in the Welch test makes the test too conservative. A reasonable compromise in this situation is to choose φ_i such that $1 \leq \varphi_i \leq c_i$.

This defines a new class of Welch type test statistics whose properties can be adjusted accordingly by choosing the control parameter, φ_i, appropriately. Our proposed test, which we shall henceforth call the *adjusted Welch test*, uses the weights $w_i^* = n_i/\varphi_i s_i^2$ in the Welch test, where $1 \leq \varphi_i \leq c_i$. That is the adjusted Welch test, S_{aw}, is given by:

$$S_{aw} = \frac{\sum_{i=1}^{K} w_i^* (\bar{y}_{i.} - \sum_{j=1}^{K} h_j^* \bar{y}_{j.})^2}{\left((K-1) + 2 \cdot \frac{K-2}{K+1} \cdot \sum_{i=1}^{K} \frac{1}{n_i - 1}(1 - h_i^*)^2\right)}, \tag{1.11}$$

where $h_i^* = w_i^* / \sum_{i=1}^{K} w_i^*$, $i = 1, \ldots, K$.
Under H_0, the adjusted Welch statistic, S_{aw}, is distributed approximately as an F-variable with $K-1$ and ν_g^* degrees of freedom, with

$$\nu_g^* = \frac{(K^2 - 1)/3}{\sum_{i=1}^{K} \frac{1}{n_i - 1}(1 - h_i^*)^2}.$$

Test: Reject H_0 at α level if $S_{aw} > F_{K-1, \nu_g^*; 1-\alpha}$.

When the sample sizes are large, S_{aw} approaches the Welch test, that is, $(n_i + \delta_1)/(n_i + \delta_2) \overset{n_i \to \infty}{\longrightarrow} 1$. With small sample sizes, our statistic will help correct the liberality witnessed in the Welch test.

To assess the relative performance of these test statistics in terms of the actual levels of significance attained, we will consider levels between 4% and 6% to be satisfactory, that is, following Cochran's rule of thumb (cf. Cochran, 1954).

1.3 SIMULATION STUDY AND DISCUSSION

In order to see the effect of balancedness and unbalancedness, as well as variance homogeneity and heterogeneity, a simulation study was conducted with sampling experiments determined by the number of studies, sample sizes and the variances in each study. In the first sampling experiment the following patterns and combinations of the number of studies, sample sizes and variances were considered (cf. Tables 1.1, 1.2, 1.3, and 1.4): Balanced samples and homogeneous variances, unbalanced samples combined with homogeneous variances. The next experiment investigated the effect of variance heterogeneity on the empirical Type I error rates. We matched balanced and unbalanced sample sizes with heterogeneous variances. In the unbalanced sample size cases, large sample sizes were separately paired with small and large variances. To investigate the effect of a large number of studies, we started with $K = 3$ studies and made independent replications to give $K = 6, 2 \times (.)$, $K = 9$, $3 \times (.)$, and $K = 18, 6 \times (.)$. We will use the term small sample to refer to $n_i = 5$, and moderate for $n_i = 10$, 15, $i = 1, \dots, K$. However, if any of the sample sizes, n_i, is greater or equal to 20, then the constellation will be taken to be of large samples.

Table 1.1 reports the actual significance levels for $K = 3$, Table 1.2 for $K = 6$, Table 1.3 for $K = 9$ and Table 1.4 for $K = 18$. For the adjusted Welch test, S_{aw}, we have taken $\varphi_i = (n_i + 2)/(n_i + 1), i = 1, \dots, K$. From these Tables, we make the following observations in order of the various tests presented in Section 1.2 above:

a) The ANOVA *F* Test

In the case when the number of populations, $K = 3$:

i. for balanced samples sizes and homoscedastic cases, the test, as expected, keeps the nominal level;

ii. for balanced and heterogeneous variance cases, the test keeps control of the significance level. This trend is maintained with increasing sample sizes;

iii. for unbalanced and homoscedastic cases, the test keeps the nominal level;

iv. for the unbalanced and heterogeneous cases, if small samples are matched with small variances, the test tends to be conservative. However, when small sample sizes are paired with large variances, the test becomes liberal. This pattern remains largely unchanged even if the sample sizes are increased.

For $K = 6$, $K = 9$ and $K = 18$, the observations made in i. to iv. above still hold; except for balanced designs and heterogeneous variances where the test becomes more liberal with increasing number of populations.

Table 1.1 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 3$

Sample Sizes	Variances	$\hat{\alpha}\%$						
(n_1, n_2, n_3)	$(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.0	4.8	12.2	4.1	3.8	3.8	3.3
	(1,3,5)	6.0	5.0	13.5	4.6	4.2	4.2	3.6
(10,10,10)	(4,4,4)	5.1	4.9	8.4	4.9	4.6	4.6	3.9
	(1,3,5)	5.7	4.7	8.2	5.1	4.5	4.5	3.9
(20,20,20)	(4,4,4)	5.1	4.9	6.5	5.0	4.9	4.9	4.2
	(1,3,5)	5.6	4.8	6.4	5.4	4.7	4.7	4.2
(40,40,40)	(4,4,4)	4.9	4.9	5.6	4.8	4.8	4.8	4.5
	(1,3,5)	5.9	5.2	5.8	5.8	5.0	5.0	4.8
(5,10,15)	(4,4,4)	5.0	5.3	10.2	5.1	4.8	5.4	4.2
	(1,3,5)	2.4	4.9	8.9	5.6	4.7	4.5	3.8
	(5,3,1)	12.3	5.4	11.5	5.3	5.0	6.2	4.4
(10,20,30)	(4,4,4)	5.2	5.3	7.7	5.1	4.9	5.3	4.5
	(1,3,5)	2.2	4.9	6.5	5.5	4.6	4.5	4.2
	(5,3,1)	12.9	5.5	8.1	5.6	5.2	5.9	4.5
(20,40,60)	(4,4,4)	4.8	4.9	5.9	4.9	4.7	4.9	4.4
	(1,3,5)	2.1	5.1	5.8	5.7	4.7	4.6	4.5
	(5,3,1)	12.5	4.9	6.4	5.5	5.0	5.4	4.4

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.2 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 6$

Sample Sizes	Variances	$\hat{\alpha}\%$						
$2\times$	$2\times$							
(n_1, n_2, n_3)	$(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.2	6.2	22.1	4.1	3.3	3.3	4.1
	(1,3,5)	6.6	6.1	22.4	4.8	3.7	3.7	4.3
(10,10,10)	(4,4,4)	5.1	5.1	11.4	4.8	4.2	4.2	3.7
	(1,3,5)	6.3	5.2	12.0	5.6	4.3	4.3	3.7
(20,20,20)	(4,4,4)	4.8	4.7	7.7	4.7	4.3	4.3	3.8
	(1,3,5)	6.0	4.8	7.7	5.7	4.4	4.4	4.0
(40,40,40)	(4,4,4)	4.7	4.6	6.0	4.6	4.4	4.4	4.2
	(1,3,5)	6.8	5.4	6.9	6.6	5.0	5.0	4.9
(5,10,15)	(4,4,4)	5.0	6.3	15.5	4.7	4.0	4.5	4.7
	(1,3,5)	2.4	5.5	13.1	5.9	4.3	4.2	3.8
	(5,3,1)	16.3	6.7	16.7	5.7	4.6	5.5	5.0
(10,20,30)	(4,4,4)	5.5	5.7	9.7	5.2	4.7	4.9	4.8
	(1,3,5)	2.3	5.2	8.3	6.5	4.8	4.7	4.2
	(5,3,1)	16.3	5.7	10.2	6.3	4.8	5.5	4.7
(20,40,60)	(4,4,4)	5.2	5.3	7.2	5.2	4.8	5.0	4.6
	(1,3,5)	2.6	5.5	7.1	6.7	5.1	5.0	4.7
	(5,3,1)	15.3	4.8	6.7	6.3	4.9	5.2	4.1

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.3 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 9$

Sample Sizes	Variances	$\hat{\alpha}\%$						
$3\times$ (n_1, n_2, n_3)	$3\times$ $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.3	7.3	28.6	4.3	3.2	3.2	4.7
	(1,3,5)	6.5	7.8	28.7	4.7	3.3	3.3	5.1
(10,10,10)	(4,4,4)	5.1	6.2	14.8	4.9	4.0	4.0	4.3
	(1,3,5)	7.0	6.0	14.5	6.2	4.5	4.5	4.3
(20,20,20)	(4,4,4)	5.2	5.4	9.1	5.1	4.6	4.6	4.4
	(1,3,5)	6.6	5.1	9.1	6.3	4.5	4.5	4.2
(40,40,40)	(4,4,4)	5.0	5.2	7.0	5.0	4.7	4.7	4.5
	(1,3,5)	6.6	4.9	6.9	6.5	4.8	4.8	4.4
(5,10,15)	(4,4,4)	5.3	7.0	19.3	4.9	4.1	4.5	4.9
	(1,3,5)	2.2	6.6	16.9	6.2	4.1	4.0	4.6
	(5,3,1)	18.7	7.6	20.9	5.5	4.1	5.1	5.5
(10,20,30)	(4,4,4)	4.9	5.5	10.7	4.8	4.3	4.4	4.1
	(1,3,5)	2.1	5.2	9.6	6.4	4.6	4.5	4.0
	(5,3,1)	17.6	5.9	10.7	5.8	4.3	4.8	4.6
(20,40,60)	(4,4,4)	5.2	5.5	8.0	5.3	5.1	5.2	4.9
	(1,3,5)	2.3	5.4	7.3	7.0	5.2	5.1	4.0
	(5,3,1)	18.1	5.3	7.6	6.4	4.8	4.9	4.5

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.4 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 18$

Sample Sizes	Variances	$\hat{\alpha}\%$						
6× (n_1, n_2, n_3)	6× $(\sigma_1^2, \sigma_3^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	4.9	11.7	46.3	3.8	2.5	2.5	7.1
	(1,3,5)	7.1	12.4	46.7	4.1	3.2	3.2	7.8
(10,10,10)	(4,4,4)	5.3	7.0	20.9	5.1	4.0	4.0	4.4
	(1,3,5)	7.0	6.8	21.1	6.2	3.8	3.8	4.2
(20,20,20)	(4,4,4)	4.8	5.2	11.3	4.8	4.2	4.2	4.1
	(1,3,5)	7.0	5.2	11.0	6.7	4.7	4.7	4.0
(40,40,40)	(4,4,4)	4.8	5.3	7.9	4.8	4.6	4.6	4.4
	(1,3,5)	7.1	5.3	8.0	6.9	5.0	5.0	4.4
(5,10,15)	(4,4,4)	5.2	9.7	28.6	4.8	3.5	4.0	6.4
	(1,3,5)	1.7	8.2	26.6	6.8	4.5	4.4	5.1
	(5,3,1)	24.9	10.2	30.0	5.7	3.7	4.6	7.1
(10,20,30)	(4,4,4)	5.3	6.2	14.0	5.3	4.5	4.7	4.4
	(1,3,5)	1.5	5.9	13.0	7.0	4.5	4.4	4.2
	(5,3,1)	24.1	6.1	14.4	6.5	4.4	4.8	4.2
(20,40,60)	(4,4,4)	4.8	5.2	8.5	4.8	4.3	4.4	4.0
	(1,3,5)	1.5	5.3	8.5	6.5	4.4	4.4	4.5
	(5,3,1)	23.2	5.0	8.2	6.3	4.3	4.5	4.0

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

b) The Welch Test

For $K = 3$:

i. for balanced sample sizes and homoscedastic cases, the test keeps the required nominal level. Increasing the individual sample sizes has no significant effect on the attained levels of significance;

ii. for balanced and heterogeneous variances, the test keeps the nominal level and this is not significantly affected by enlarging the sample sizes;

iii. for unbalanced and homogeneous variance case, the test performs well and there is no significant effect in increasing the sample sizes;

iv. for unbalanced and heterogeneous variance cases, the test attains acceptable significance levels both for small and large sample sizes.

For $K = 6$:

i. for balanced sample sizes and homoscedastic cases, the test keeps the required nominal level but seems to be a bit liberal when the sample sizes are small. Increasing the individual sample sizes has the effect of making the attained levels of significance closer to the nominal level;

ii. for balanced and heterogeneous variances, the test keeps the nominal level but, is a bit more liberal when the sample sizes are small;

iii. for unbalanced and homogeneous variance cases, if one of the sample sizes is small, the test becomes liberal otherwise, it attains acceptable levels of significance;

iv. The test attains acceptable significance levels for unbalanced and heteroscedastic cases, except when one of the sample sizes is small and is paired with a large variance.

For $K = 9$:

i. for balanced samples sizes and homoscedastic cases, the Welch test is liberal for small samples but works quite well for moderate to large sample sizes;

ii. for balanced and heterogeneous variances, the test is liberal for small variance cases but, works well for moderate to large samples;

iii. for unbalanced and homogeneous variance cases, for small sample sizes the test becomes liberal but, in general it attains acceptable levels of significance;

iv. in the case of unbalanced samples and heterogeneous variances, the test is slightly liberal when one of the sample sizes is small. Otherwise, increasing the sample sizes makes the test better.

For $K = 18$:

i. for balanced samples and equal variances in all the groups, the test is liberal for small and moderate samples. For large samples, the test keeps good control of the significance level;

ii. for equal sample sizes in all the groups and heterogeneous variances, the test is liberal for small and moderate samples but, attains acceptable significance levels for large sample sizes;

iii. for unbalanced and homogeneous variances, the test attains acceptable significance levels only when all the sample sizes are above or equal to 20. Otherwise, the test is liberal;

iv. for unbalanced samples and unequal variances in the groups, the test controls the nominal level when all the sample sizes are equal or greater than 20. When one of the samples is of size 10 or less, the test is more liberal when relatively large variances are combined with relatively small sample sizes.

c) Cochran's Test

For $K = 3$:

i. for balanced sample sizes and homogeneous variances, the test is largely liberal. This liberality is much more pronounced for very small samples and reduces drastically for moderate to large samples;

ii. for balanced sample sizes and heterogeneous variances, the test attains levels which are above the acceptable upper limit but, there is some improvement as the sample sizes increase;

iii. for unbalanced sample sizes and homogeneous variances, liberality is clear but reduces with increased sample sizes;

iv. for unbalanced samples and heterogeneous variances, the test is generally liberal but, if small variances are paired with small sample sizes, the attained levels are relatively lower than in case of large variances paired with small sample sizes. Increasing the sample sizes improves the attained significance levels.

For $K = 6$:

i. for homogeneous variances and balanced samples, the liberality of the test persists but it is interesting to note, for example, that when the all the sample sizes are $n_i = 5, i = 1, \ldots, K$, the level attained is 22.1%. This improves appreciably by about a half to 11.4% when the sample sizes are doubled, $n_i = 10, i = 1, \ldots, K$;

ii. for balanced samples and heterogeneous variances, the improvement in the attained levels similar to i) is observed; but the test remains liberal;

iii. for unbalanced samples and homogeneous variances, the test is liberal and improves with increased sample sizes;

iv. for unbalanced samples and heterogeneous variances, the test is liberal and, as expected, attains better levels as the sample sizes increase.

For $K = 9$:

i. for homogeneous variances and balanced samples, the test attains unacceptable significance levels for very small sample sizes, for example, 28.6% for all $n_i = 5$. This scenario dramatically improves to 14.8% when all $n_i = 10$ and 7.0% when $n_i = 40$;

ii. for balanced and heterogeneous variances, at very small samples the levels attained are unacceptable. This improves as the sample sizes increase. However, the test still remains liberal;

iii. for unbalanced samples and homogeneous variances, even though the test is liberal, an improvement of the attained levels with increased samples is clear;

iv. for unbalanced samples and heterogeneous variances, the test is liberal and improves as the samples sizes increase. Note, that when small samples are combined with small variances the attained significance levels are relatively better than when small samples are combined with large variances.

For $K = 18$:

For all cases of balancedness and unbalancedness, and homogeneity and heterogeneity, the test attains unacceptable significance levels. Notice the extreme liberality when the sample sizes are small.

d) Brown-Forsythe (B-F) Test

For $K = 3$:

For balanced samples and homoscedastic cases, the test attains acceptable significance levels which remain significantly unaffected by increases in individual sample sizes. This observation is largely true for balanced sample sizes combined with heterogeneous variances, unbalanced sample sizes combined with homogeneous variances and unbalanced samples combined with heterogeneous variances.

For K=6:

i. for balanced samples and homogeneous variances, the test attains acceptable levels for all sample sizes;

ii. for balanced samples and heterogeneous variances, the test keeps significance levels for small samples but, for large samples, the test becomes liberal;

iii. for unbalanced samples and homogeneous variances, the levels attained by this test are acceptable;

iv. for unbalanced samples and heterogeneous variances, the test attains acceptable levels for small samples. In the other cases the test becomes liberal.

For $K = 9$:

i. for balanced sample sizes and homogeneous variances, the test attains levels which are within the acceptable range;

ii. for balanced samples and heterogeneous variances, the test becomes liberal for large sample sizes;

iii. for unbalanced samples and homoscedastic cases, the test attains good levels;

iv. for unbalanced samples and heterogeneous variances, the test is in general liberal, except for the case when a small sample is paired with a large variance.

For $K = 18$:

i. for equal sizes in the groups and homogeneous variances, the test attains an acceptable significance level except when the sample sizes are small, in which case it tends to be a bit conservative;

ii. for equal sample sizes and heterogeneous variances, the test attains acceptable levels only when the samples are small. Otherwise, the test becomes liberal even for large sample sizes;

iii. for unbalanced samples and equal variances, the test keeps the significance level within acceptable limits. This is true for small to large sample cases;

iv. for unbalanced samples and unequal variances, the test is always liberal,with the liberality being more pronounced when relatively small samples are paired with relatively small variances.

e) The Modified Brown-Forsythe Test

For $K = 3$:

for balanced or unbalanced sample sizes matched with homogeneous or heterogeneous variances, the levels attained are acceptable, except for one case where the test is a bit conservative.

For $K = 6$:

i. for balanced samples and homogeneous variances, the test is a bit conservative for small samples. Otherwise, with moderate to large sample sizes, the test attains acceptable levels. This is also true for balanced samples combined with heterogeneous variances;

ii. for unbalanced samples matched with homogeneous or heterogeneous variances, the test attains acceptable significance levels.

For $K = 9$ and $K = 18$, all the observations for $K = 6$ above are true but, the degree of conservativeness in cases of small samples increases with increasing K.

f) The Approximate ANOVA *F* Test

This test is very similar to the modified Brown-Forsythe test given above. As noted in Section 1.2 above, when the sample sizes in the groups are equal, the results are expected to be the same. Checking Tables 1.1-1.4, we see that the attained levels are all equal for the two tests for balanced designs. Consequently, for $K = 3$, 6, 9 and 18, the observations made above for the modified Brown-Forsythe test also hold for this test (approximate ANOVA *F* test).

For unbalanced samples and $K = 3$, 6, 9 and 18, the test attains levels which are very close to the Brown-Forsythe test for all values of K, save for small differences. For example, in one case, $K = 3$, when the sample combination is (5,10,15) and the corresponding variances (5,3,1), the approximate ANOVA *F* test is liberal (6.2%) whereas the modified Brown-Forsythe attains an acceptable level. Further, even though both tests attain acceptable significance levels, the approximate ANOVA test has relatively higher levels when relatively large variances are paired with relatively small sample sizes and when the variances in the groups are all equal. However, when relatively small variances are combined with small sample sizes, the approximate ANOVA *F* test attains relatively lower significance levels compared to the modified Brown-Forsythe test.

g) The Adjusted Welch Test

For $K = 3$:

i. for balanced samples and homogeneous variances, the test keeps control of the nominal significance level only when the sample sizes are large otherwise, the test is conservative;
ii. for balanced samples and heterogeneous variances, the test attains acceptable significance level only when the sample sizes are large otherwise, the test is conservative;
iii. for unbalanced samples and homogeneous variances, the adjusted Welch test attains acceptable levels;
iv. for unbalanced samples and heterogeneous variances, the actual levels are within the acceptable limits.

For $K = 6$:

i. when the sample sizes and error variances in the groups are equal, the test does not seem to keep good control of the nominal significance level;
ii. for equal sample sizes and heterogeneous variances, the test attains reliable significance levels only when all the sample sizes are equal to 40;
iii. for unbalanced samples and homogeneous variances, the test attains acceptable significance levels;

iv. for unbalanced samples and heterogeneous variances, the actual levels attained are within the limits (4%, 6%).

For $K = 9$:

For all cases of balancedness, unbalancedness, homoscedasticity, and heteroscedasticity, the actual levels attained are acceptable.

For K=18:

For balanced and small sample sizes the test is liberal. For moderate to large sample sizes the test attains acceptable levels. In the case of unbalancedness, when one sample is small the test is liberal except when a small sample size is paired with a small error variance. In this case the test attains its level. For moderate to large samples, the levels attained are in the acceptable range.

In general, we see that for the Welch test, increasing the number of populations has no significant effect on the levels attained, except that for balanced small samples the test becomes too liberal. In this case the adjusted Welch test is preferred. For the Cochran test, increasing the number of populations has the effect of dramatically inflating the significance levels when the sample sizes are small and sometimes also for moderate samples. For the Brown-Forsythe test, increasing the number of studies does not significantly affect the levels attained by the test. For the modified Brown-Forsythe test, the attained significance levels are all within the acceptable range for $K = 3$, 6, 9, and 18, save sometimes for small samples where the test becomes more conservative with an increasing number of populations. This behavior is also true for the approximate ANOVA F test. The modified Brown-Forsythe test rarely attains significance levels above 5%. This is true regardless of the number of populations.

1.4 CONCLUSION

The modified Brown-Forsythe and the approximate ANOVA F test are relatively least affected by changes in the sample sizes and number of populations except when the number of groups is large and the corresponding sample sizes are small, in which case these tests become too conservative. The Brown-Forsythe test should not be used when the number of populations, K, is large with large sample sizes and heterogeneous variances. We will recommend the test for small individual samples regardless of the number of populations. The Welch test can be recommended in the case of heterogeneous variances, except when the sample sizes are small and the number of studies is large. In this case we recommend a suitable adjustment of the adjusted Welch test. This test (adjusted Welch test) has the advantage that the weights, $w_i^*, i = 1, \ldots, K$, have a control parameter, $\varphi_i, i = 1, \ldots, K$, which can be adjusted accordingly.

In case of homoscedasticity regardless of balanced or unbalanced designs, the ANOVA F test is the optimal test (cf. Lehmann, 1986) and the observed deviations from the nominal level are due to the simulation experiment.

REFERENCES

Asiribo, O., & Gurland, J. (1990). Coping with variance heterogeneity. *Communications in Statististics – Theory and Methods*, *19*, 4029–4048.

Böckenhoff, A., & Hartung, J. (1998). Some corrections of the significance level in meta-analysis. *Biometrical Journal*, *40*, 937–947.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, *Supplement 4*, 102–118.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2-tests. *Biometrics*, *10*, 417–451.

De Beuckelaer, A. (1996). A closer examination on some parametric alternatives to the ANOVA F-test. *Statistical Papers*, *37*, 291–305.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.

Mehrotra, D. V. (1997). Improving the Brown–Forsythe solution to the generalized Behrens–Fisher problem. *Communications in Statistics – Simulation and Computation*, *26*, 1139–1145.

Patel, J. K., Kapadia, C. P., & Owen, D. B. (1976). *Handbook of statistical distributions.* New York: Marcel Dekker.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

2

Differences in the Results of Two Meta-Analytical Approaches

Ralf Schulze
Heinz Holling
Heiko Großmann
Andreas Jütting
Michaela Brocke
Psychologisches Institut IV
Westfälische Wilhelms-Universität Münster

Summary

Two meta-analytical approaches for the analysis of correlation coefficients as effect sizes are distinguished. The approaches differ mainly with respect to the effect size being aggregated (r vs. Fisher's z), the weights used in aggregation, estimators for the standard error of the aggregate, and computational procedure for the homogeneity test. The performance of the approaches is compared with regard to bias of estimators, coverage rates of confidence intervals, and Type I error of the homogeneity tests. To comparatively evaluate the approaches, a simulation study with a varying number of studies, number of subjects per study, and population correlations was conducted. The situations in the simulation study are restricted to homogeneous cases. The results show that, overall, the approach as proposed by Hedges and Olkin (1985) as well as Rosenthal (1991) is preferable to the alternative approach of Hunter and Schmidt (1990) for the situations under study.

2.1 INTRODUCTION

As almost any other statistical method, meta-analysis has its own history of developments. Many of its procedural details emerged from adaptations of the method to specific problems of application. This can easily be observed by comparing the major book publications by the various early protagonists of the method in psychology (Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1991)[1].

The development of meta-analysis was at least partly motivated in the late 1970s and early 1980s by a widespread dissatisfaction with the state of the social sciences and psychology in particular (see Hunter & Schmidt, 1990). The situation was characterized by a large number of published studies on various subject matters apparently showing heterogeneous results. There were only few areas of research for which clear conclusions about the effectiveness of interventions or the quality of models to explain and predict human behavior could be drawn from the literature. Facing this state of affairs, researchers from educational, clinical, and industrial/organizational (I/O) psychology began to develop methods to systematically integrate research findings across studies to overcome deficiencies associated with more narrative methods of literature reviews. Interestingly, these developments were done in parallel in subdisciplines of psychology.

Glass and coworkers were the first to publish a comprehensive treatment of the topic (Glass et al., 1981) with a focus on the evaluation of educational and clinical research questions. Accordingly, their main interest was the development of methods for cumulating results from experimental designs. The most prevalent effect sizes in this are of research were therefore (standardized) mean differences between groups.

In contrast, one of the focal research questions in I/O psychology, and personnel selection in particular, has been the question of whether the validities of personnel selection procedures are situation specific or can be generalized across situations. The validities were ordinarily assessed by the correlations of results from procedures to select applicants with criterion measures like supervisory ratings, for example. Thus, the main concern here was to develop procedures to accumulate the effect size r from a series of studies (Hunter et al., 1982).

In addition to differences in emphasis on effect size families, there also emerged a plethora of further differences between meta-analytical methods, like the introduction of the 75% rule for the detection of heterogeneity in the approach by Hunter and Schmidt (1990), which is unique to their procedures. Furthermore, other researchers developed methods to aggregate study results like p-values or other outcomes of significance tests (Rosenthal & Rubin, 1979), for example, and summarized their developments in their own treatment of

[1] The history of meta-analysis does not begin with these references and is much older, as Hunt (1997) and Olkin (1990) have pointed out.

the topic (Rosenthal, 1991). Additionally, some researchers focused more on the statistical steps of meta-analysis (Hedges & Olkin, 1985).

As a result of these attempts to establish a comprehensive and elaborate set of methods and procedures for the purpose of integrating research findings, there emerged distinguishable *approaches* to meta-analysis. These approaches differ with respect to a series of attributes and are associated with different areas of application, at least in psychological research. Despite attempts to systemize approaches (e. g., Bangert-Drowns, 1986) along characteristics like units or outcomes of analysis, for example, the approaches as described above still seem to prevail in different subdisciplines of psychology.

As a new and largely statistical method, meta-analysis diffused astonishingly fast into psychological research practice and was quickly adopted by the research community. This gave rise to a rapid growth of number of articles that used meta-analysis instead of narrative reviews to summarize the state of the art on a research question. It is noteworthy in this context that in applications of meta-analytical methods, researchers almost exclusively used the respective approach of their field. The approach by Hunter and Schmidt, for example, strongly dominated in I/O psychology.

The story of meta-analysis differs between psychology and other disciplines like medicine, where researchers were more reluctant to use meta-analysis (for a critical assessment of the method, see Feinstein, 1995, for example). Comprehensive treatments of meta-analytic methods (e.g., Sutton, Abrams, Jones, Sheldon, & Song, 2000) also became available in medicine much later than in psychology. Furthermore, the focus of expositions in psychology and medicine differs with respect to effect sizes of main interest, specialized techniques, and many other attributes.

Thus, although there is a general purpose of application common to all methods of meta-analysis, a large number of procedures and techniques exist. This supports the view that meta-analysis should not be regarded as a single method but as a conglomerate of methods to integrate research findings encompassing statistical as well as non-statistical steps. Despite existing differences between approaches, there are also efforts to point out a general structure of the statistical procedures to aggregate effect sizes (Shadish & Haddock, 1994). But even when this general structure can be regarded as accepted, there still remain more subtle differences between approaches. Such differences might influence the meta-analytic results and therefore also the substantive conclusions drawn from these results.

In this chapter, we focus on approaches of meta-analysis developed in psychological research that are designed for the aggregation of the correlation coefficient as an effect size. As a consequence, procedures for the aggregation of standardized mean differences or other effect sizes will not be of concern here. The specific statistical procedures of the relevant approaches will first be presented in the next section. After an analysis of their properties and accentuation of theoretical differences, the results of a simulation study will be presented in the subsequent section. The aim is to make a comparative evaluation of the approaches under scrutiny with respect to their statistical performance.

2.2 COMMON META-ANALYTICAL APPROACHES IN THE SOCIAL SCIENCES

As has been described in the preceding section, there are several approaches of meta-analysis in psychology that differ with respect to a large number of attributes. Of these, the procedures developed in the context of educational research by Rosenthal (1991), in I/O psychology by Hunter et al. (1982), and with a more general statistical focus by Hedges and Olkin (1985) are of main concern here. Because the methods summarized in the book by Rosenthal (1991) were preceded by several journal publications in collaboration with Rubin (Rosenthal & Rubin, 1979, 1982), this approach will be labelled as Rosenthal-Rubin (RR) approach. For applications in I/O psychology, Hunter and Schmidt published a successor of their first book publication (Hunter & Schmidt, 1990), which has become the main reference in this field of research. Their approach will therefore conveniently be labelled as Hunter-Schmidt (HS) approach. Finally, the meta-analytical procedures summarized in the book by Hedges and Olkin (1985) will be abbreviated in the following as HO approach.

The three approaches cannot be comprehensively evaluated here in all of the steps they propose for meta-analysis, partly because they are not equally specific. We therefore focus on the step of statistical aggregation of effect sizes (correlations) to arrive at an estimate of the mean effect size, estimates for confidence limits for the mean effect size, and the homogeneity test. These steps are element of almost every published meta-analysis in the social sciences but represent only a core of the statistical procedures. The elaborate artifact corrections, for example, proposed and advocated by Hunter and Schmidt (1990) are not considered here because other approaches do not specify alternative procedures or do not recommend using corrections for unreliability of measures (Rosenthal, 1991).

The three approaches were also distinguished in a previous comparison of these methods (Johnson, Mullen, & Salas, 1995), which we took as a starting point for our evaluation. However, Johnson et al. specified the approaches in a form that differs from the specification presented in detail below.

One major difference to the specification of Johnson et al. (1995) is that the approaches of HO and RR are not distinguished here. The reasons for this are twofold. First, the HO approach is specified by Johnson et al. (1995) as using the d-statistic as effect size. With correlation coefficients as effect sizes, this would require to convert the correlations to standardized mean differences before aggregation by using an appropriate transformation. This is exactly what Johnson et al. have done in their evaluation of the approaches. Unfortunately, we cannot see any reason why one would in general want to convert a database consisting entirely of r to d. More importantly, we cannot see any indication in the work of Hedges and Olkin for a recommendation to do that. Instead, Hedges and Olkin (1985) present specific formulas for correlations in meta-analysis which we will use in our simulation study. Second, Johnson et al. presented the mean effect size estimator for the RR approach to use study sample sizes (N_i) as weights. We note, however, $N_i - 3$ being recommended by Rosenthal (1991) as weights, the same as in the HO approach for correlation coefficients. This also has an effect on the standard error for the mean effect size estimator which becomes the same as in the HO approach. Furthermore, although Rosenthal and Rubin (1979) have indeed presented procedures to summarize significance levels, they do not strictly advocate using these methods for the case of interest here. By consulting Rosenthal's work (Rosenthal, 1991) it becomes evident that for the present purposes the approaches by HO and RR are in fact identical. Thus, in contrast to Johnson et al. (1995) we do not distinguish them in the following.

We next turn to a specification of the remaining two approaches. Differences between them can best be explained by considering the basic formulas shown in Table 2.1.

Table 2.1 Basic Formulas for the HO/RR- and HS-Approach (Homogeneous Case)

	HO/RR	HS
Estimator for MES	$\bar{z} = \frac{\sum_{i=1}^{k} \left(\hat{\sigma}_{z_i}^{-2}\right) z_i}{\sum_{i=1}^{k} \hat{\sigma}_{z_i}^{-2}}$	$\bar{r} = \frac{\sum_{i=1}^{k} N_i r_i}{\sum_{i=1}^{k} N_i}$
Variance of MES	$\hat{\sigma}_{\bar{z}}^{2} = \left(\sum_{i=1}^{k} N_i - 3k\right)^{-1}$	$\hat{\sigma}_{\bar{r}}^{2} = \frac{1}{k}\left(\frac{\sum_{i=1}^{k} N_i \left(r_i - \bar{r}\right)^2}{\sum_{i=1}^{k} N_i}\right)$
Homogeneity test	$Q = \sum_{i=1}^{k} \left(N_i - 3\right)\left(z_i - \bar{z}\right)^2$	$Q = \frac{\sum_{i=1}^{k} \left(N_i - 1\right)\left(r_i - \bar{r}\right)^2}{\left(1 - \bar{r}^2\right)^2}$

Note. HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach, MES = Mean Effect Size.

The HO/RR and HS approach can both be used to summarize a database consisting of a total of k correlations. Whereas by using the HS approach the untransformed correlations r are taken to arrive at an estimate for the mean effect size (MES), the correlations have to be transformed in the HO/RR approach by applying the following transformation

$$z_i = .5 \times \ln\left(\frac{1+r_i}{1-r_i}\right) = \tanh^{-1}(r_i) \tag{2.1}$$

The transformation given in Equation 2.1 was first introduced by Fisher (1915) in the context of deriving the sampling distribution of the correlation coefficient and is mostly labelled as Fisher's z. It is important to note that an estimator for the approximate sampling variance of the transformed correlations z_i is given by $\hat{\sigma}^2_{z_i} = 1/\left(N_i - 3\right)$. Thus, the standard error for the transformed correlations *only* depends on the sample size N_i. Whereas the standard error of the correlation coefficient depends on sample size and the population correlation ρ, Fisher's z stabilizes the variance in the sense that it is independent of ρ. By inspecting the formula for the estimator of the mean effect size in the HO/RR approach in Table 2.1, it can be seen that the inverse variances of the estimator ($\hat{\sigma}^{-2}_{z_i}$) are used as weights in the aggregation process. These weights are optimal in the sense that they minimize the variance of the estimator of the MES (Hedges & Olkin, 1985). To compute a mean correlation coefficient for a set of studies, the inverse transformation given by

$$r = \frac{\exp(2z) - 1}{\exp(2z) + 1}. \tag{2.2}$$

is usually applied to $\bar{z}$.

The variances given in the second row of Table 2.1 can be used to conduct a significance test for the MES and also to construct confidence intervals. The formulas differ between the approaches because of the aforementioned use of Fisher's z in the HO/RR approach and untransformed correlations in the HS approach. Additionally, the estimator for the variance in the HS approach differs from the one specified by Johnson et al. (1995). As Schmidt and Hunter (1999) have pointed out, Johnson et al. used a wrong formula for that purpose. By using an incorrect estimator for the standard error, most of their results on the differences between approaches were invalidated.

Confidence intervals for the MES can be constructed for both approaches by using the information given in the first two rows of Table 2.1, assuming a normal sampling distribution for the MES, and applying standard procedures. Whereas on the basis of theoretical results (see Fisher, 1915; Hotelling, 1953) the normal distribution can safely be assumed in the HO/RR approach, the distribution of the correlation coefficient is known to be non-normal for nonzero population correlations ρ and small to moderate N_i. Therefore, problems may result in the HS approach for confidence intervals and statistical testing of the MES, especially when N_i and k are small and ρ is large. In our evaluation of the approaches we follow general (Wilkinson & Task Force on Statistical Inference, 1999) and specific (Schmidt & Hunter, 1995) recommendations to focus on confidence intervals and not null hypothesis testing.

In the last row of Table 2.1, the formulas for conducting a homogeneity test in both approaches are given. Although both formulas are apparently different, they follow the same structure in that the squared differences of the (transformed) effect sizes are weighted by the inverse of their variance and summed over k studies. The result is a statistic ordinarily designated Q that follows a χ^2 distribution with $k - 1$ degrees of freedom in the homogeneous case (see also Chapter 10 by Böhning & Dammann as well as Chapter 1 by Hartung, Argaç, & Makambi, this volume).

A still open question is how the approaches HO/RR and HS should be classified with respect to random vs. fixed effects models of meta-analysis. The HO/RR approach as presented in this chapter and most often used in practice, clearly represents a fixed effects model and is classified as such by its authors (Hedges & Olkin, 1985). One indication of the fixed effects model is that there is no variance component being estimated and used in the formula for the variance of the MES (see Table 2.1). In the random effects procedures specified by Hedges and Olkin (1985) a variance component is estimated and used to compute the weights applied in aggregation. The classification of the HS approach is not as simple as for the HO/RR approach. The authors are inconsistent in their own classification by stating that their methods use the fixed effects model (Hunter & Schmidt, 1990, p. 405) but also that the methods are random effects models (Hunter & Schmidt, 2000, p. 275). Furthermore, other authors also do not seem to agree (cf. Erez, Bloom, & Wells, 1996; Field, 2001; Hedges & Olkin, 1985). We note that for the classification it is essential to assess the assumptions of an approach about population parameters, whether they are constant (fixed) or possibly variable (random). Although not obvious from the procedures of the HS approach as presented in this chapter, it is an integral part of the general HS approach that – for several reasons – population correlations can be variable and are best considered as a random variable. However, Hunter and Schmidt (1990) do not present separate procedures for the different models as Hedges and Olkin (1985) do. Instead, the statistical procedures as shown in Table 2.1 for the HS approach can be applied in homogeneous as well as heterogeneous situations, that is, when the fixed or the random effects model, respectively, is appropriate. Of particular interest in this context is the variance of the MES as shown in the right column of Table 2.1. This is the formula generally recommended in the HS approach (Schmidt & Hunter, 1999) because it is supposed to hold for the heterogeneous case and "serves equally well when study effect sizes are homogeneous" (Osburn & Callender, 1992, p. 116).

To summarize, with respect to the statistical procedures of the approaches that are appropriate for the homogeneous case as presented in this chapter, there are two main differences that lead to further differences in details of the procedures. First, in the HO/RR approach Fisher's z is used whereas in the HS approach untransformed correlations are aggregated. Second, in the HO/RR approach the inverses of the (estimated) variances are applied as weights in the aggregation process whereas in the HS approach the sample sizes are used.

In the following simulation study we will comparatively evaluate the performance of the approaches in only the homogeneous case, that is, when there is only one constant population correlation ρ common to all studies.

On the basis of the outlined differences of the approaches we expect the following congruences and divergences in results in homogeneous situations for different N, k, and ρ:

1. The estimates of mean effect sizes of both approaches will be biased in cases when $\rho \neq 0$. The HO/RR will show an upward bias and the results for the HS approach will be biased downwards. However, biases will in general be negligibly small, except for cases in which N is very small.
2. The absolute biases will be larger for the HO/RR approach but the absolute difference between biases of the approaches will be small.
3. The performance for the confidence intervals as assessed by the coverage of the true parameter value will be better for the HO/RR approach when N and k are small. In other cases both approaches will show similar performance.
4. The performance for the homogeneity test will be better for the HO/RR approach when N and k are small. In other cases both approaches will show similar performance.
5. We will not be able to replicate most of the results of Johnson et al. (1995) but our results will be in general agreement with those reported by Field (2001).

Predictions concerning the bias of the estimators of the MES are based on the theoretical results given in the seminal paper by Hotelling (1953). He gives estimates for biases of the transformed and untransformed correlation coefficient which can be used to deduce predictions one and two. However, we also note that previous Monte-Carlo studies on the bias of Fisher's z and r have found a smaller bias for Fisher's z which contradicts prediction two (e.g., Corey, Dunlap, & Burke, 1998; Field, 2001).

The expected superiority of the HO/RR approach for confidence intervals is based on a faster asymptotic of the distribution of the statistic ($\bar{z}$) to the normal distribution in the HO/RR approach as compared to r in the HS approach, for which convergence of the sampling distribution to the normal distribution is remarkably slow. Accordingly, the asymptotic behavior of the Q-statistic is also assumed to be better for the HO/RR procedure. Furthermore, for confidence intervals in the HS approach the mean sampling error of the correlations in studies is used, which may be influenced by what Hunter and Schmidt (1990) call "second order sampling error", that is, inaccuracies in estimation when N and/or k are small.

Finally, we note that there are two kinds of asymptotics relevant for the predictions. First, as the sample size of studies N grows larger but the number of studies k remains constant, results are expected to converge to theoretical predictions derived from large sample theory of the statistics. Second, as k grows larger but N remains constant, results for the estimators need not converge to true values being estimated. Thus, we expect larger N to have a more profound effect on the results in comparison to an increase in k.

2.3 SIMULATION STUDY

To comparatively evaluate the two approaches, a C++ program was written to perform all computations in the situations under study. The procedures to generate the database for applying the computational procedures of the approaches follow the descriptions given by Corey et al. (1998). Details of the computational procedure are reported in Schulze (in press). As already indicated, the main parameters varied in the Monte-Carlo study are number of studies k to be aggregated, number of subjects per study N, and the population correlation coefficient ρ. In the following subsection, the design of the study is described in more detail, and results are presented in the subsequent subsection.

2.3.1 Design and Procedure

The levels used for the number of studies were $k = 8$, $k = 32$, and $k = 128$. They span a wide range of k to explore the results for the approaches in a more typical case ($k = 32$) (see Cornwell, 1988) as well as extreme cases. The same is true for the number of subjects which was varied across the following levels: $N = 16$, $N = 64$, $N = 128$, and $N = 256$. For the population correlation only positive values were used because results were expected to be similar in the negative range of values. The levels of ρ used in the Monte-Carlo study were: $\rho = 0$, $\rho = .10$, $\rho = .30$, $\rho = .50$, $\rho = .70$, and , $\rho = .90$. Again, these values were chosen to explore the performance of the approaches across a wide range of values. The more typical values in psychological research are in the range from 0 to .50. However, there are also research questions in psychology for which correlations to be aggregated can be much higher as in studies on the reliability of a measurement instrument. Thus, very high values were also included in our study.

All levels of the three design features were fully crossed, that is, we distinguished a total of $3 \times 4 \times 6 = 72$ situations. Within these situations all levels were held constant in the simulation procedure. For example, for the situation $k = 8$, $N = 16$, and $\rho = 0$, the procedures of the approaches outlined in Table 2.1 were both applied to databases of 8 studies, all of which had a constant N of 16 subjects and the true correlation underlying all of the observed 8 effect sizes was zero. The computations for all of the 72 situations were repeated 10,000 times and means across these iterations were computed for the statistics of interest. By holding ρ constant within situations, we investigated the homogeneous case for which the fixed effects methods of meta-analysis are appropriate.

2.3.2 Evaluation Criteria

The performance of the approaches with respect to estimation of ρ is straightforward. We computed the estimates for both approaches in all situations and compared them to the known true values. Deviations from the true values indicate bias of the estimators.

For the confidence intervals, the number of intervals covering the population correlation ρ were counted in all iterations and divided by the number of iterations (10,000). Thus, the coverage probability was estimated by the coverage rates. This is similar to the procedure used by Brockwell and Gordon (2001). All confidence intervals were computed with a prescribed coverage probability of 95%, so that the expected coverage rate is .95 for all situations. Because high coverage rates can come at the cost of long interval widths, the mean widths were also computed by the difference of the mean upper and lower limit across all iterations. This will enable a comparison of the approaches with respect to estimated coverage probabilities when coverage rates are (almost) equal, for example. Ceteris paribus, the approach showing smaller intervals shows better performance. Additionally, estimated confidence interval widths may also be indicative for causes of potential deficiencies of coverage rates.

The homogeneity tests for both approaches and in all situations tests were conducted with $\alpha = .05$, and the rate of significant test results was assessed. Thus, we evaluated whether Type I error rates of the tests for both approaches conformed to the prescribed significance level of the tests.

2.3.3 Results

The results will be presented in separate tables for the estimates of the mean effect size, coverage rates as well as interval widths of the confidence intervals, and Type I error rates for the homogeneity tests. All tables will have the same general structure showing the results for the two approaches in two blocks of columns subdivided by levels of k. Levels of N are shown in blocks of rows where blocks represent levels of ρ.

The estimates of the mean effect sizes for the two approaches are shown in Table 2.2. There are several results for the mean effect size estimates to be highlighted. First, overall the biases are small for most combinations of ρ, k, and N. However, when N is very small (16), biases are not within rounding error for correlations. Nevertheless, we doubt that the absolute value of biases – even in the most extreme cases as observed in Table 2.2 – will affect substantial conclusions in most applications. Biases for both approaches are largest for values of ρ between .50 and .70 and are smallest for $\rho = 0$. Biases also diminish for larger values of N but do show the same behavior for increasing values of k. Thus, for neither of the approaches does adding more studies to a meta-analysis in a homogeneous situation help in obtaining less biased estimates for the population parameter, as long as study sample sizes are equal for the (added) studies.

Second, the results of the HO/RR approach show an upward bias in all situations whereas the results for the HS approach indicate underestimations of ρ. When comparing the absolute values of biases of the approaches, we note that – except for very small N – biases are equal to the third digit. Contrary

Table 2.2 Results: Estimates of Mean Effect Size

	Number of studies k					
	HO/RS			HS		
N	8	32	128	8	32	128
			$\rho = 0$			
16	.001	.000	.000	.001	.000	.000
64	.000	.000	.000	.000	.000	.000
128	.000	.000	.000	.000	.000	.000
256	.000	.000	.000	.000	.000	.000
			$\rho = .10$			
16	.102	.103	.104	.097	.097	.097
64	.101	.101	.101	.099	.099	.099
128	.101	.101	.100	.100	.100	.100
256	.100	.100	.100	.100	.100	.100
			$\rho = .30$			
16	.306	.310	.310	.290	.291	.291
64	.302	.302	.302	.299	.298	.298
128	.301	.301	.301	.299	.299	.299
256	.300	.301	.301	.299	.299	.299
			$\rho = .50$			
16	.509	.512	.513	.486	.487	.487
64	.502	.503	.503	.497	.497	.497
128	.500	.501	.501	.498	.499	.499
256	.500	.501	.501	.499	.499	.499
			$\rho = .70$			
16	.710	.712	.712	.688	.688	.687
64	.702	.703	.703	.697	.697	.697
128	.701	.701	.701	.699	.699	.699
256	.701	.701	.701	.699	.699	.699
			$\rho = .90$			
16	.905	.906	.906	.894	.894	.894
64	.901	.901	.901	.899	.899	.899
128	.901	.901	.901	.899	.899	.899
256	.900	.900	.900	.900	.900	.900

Note. N = Sample size per study, HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach.

to theoretical expectations, the method that employs Fisher's z (i.e., HO/RR) actually shows slightly less bias in cases when $N = 16$ as compared to the approach in which untransformed correlations are aggregated.

Im sum, the results for the estimates of mean effect sizes confirm prediction 1 and contradict prediction 2 on page 28. The results reported in Table 2.2 are in accordance with those reported elsewhere (Corey et al., 1998; Field, 2001; Silver & Dunlap, 1987) and show small biases in opposite directions as well as nearly equal absolute biases for both approaches.

The estimates of the coverage rates and interval widths for the two approaches are shown in Table 2.3. The values in Table 2.3 show both coverage rates of the 95% confidence intervals for the population effect size (left of slashes) and estimated widths of the intervals (right of slashes). The interval widths are given for values of r for both approaches. That is, interval limits were estimated for the Fisher's z values in the HO/RR approach, backtransformed by applying Equation 2.2 to the estimated values for the limits, and aggregated over iterations.

For the situations in which $\rho = 0$ and practically no bias was observed (see Table 2.2), a symmetrical (normal) sampling distribution can be assumed for both approaches. Because interval widths for the HO/RR approach only depend on k, N, and the quantiles of the normal distribution corresponding to the desired coverage probability, at least for $\rho = 0$ the widths should correspond very closely to theoretical expectations derived from:

$$IW = 2 \times z_{.975} \left(\sum_{i=1}^{k} N_i - 3k \right)^{-\frac{1}{2}},$$

where IW denotes interval widths and $z_{.975}$ is the .975-quantile from the standard normal distribution. The interval widths of the HO/RR approach indeed correspond exactly to the theoretical expectations, except for one case ($N = 16$, $k = 8$) showing a small difference to the expected width of $.380 - .384 = -.004$. Note, that this is also the situation, for which a very small bias was observed. The observed close correspondence to theoretical expectations lends support to the validity of the general procedure used to determine the estimates for the interval widths.

Overall, interval widths shown in Table 2.3 become smaller both for increases in k and N, as would be expected from statistical theory. The results also indicate coverage rates close to the desired level in nearly all situations for the HO/RR approach. Few exceptions are observed for combinations of large k, small N, and medium to high values of ρ. For the same combinations of values, the coverage rates of the HS approach are less than expected.

In a comparison of the performance of the approaches' procedures, two main aspects are noteworthy. First, apart from few exceptions coverage rates for the HS approach are smaller than those of the HO/RR approach and interval widths are simultaneously smaller for the HS approach. Thus, confidence intervals in the HS approach are too small and this ostensibly higher precision

Table 2.3 Results: Coverage Rates and Estimated 95% Confidence Interval Widths

	Number of studies k					
	HO/RS			HS		
N	8	32	128	8	32	128
$\rho = 0$						
16	.951/.380	.947/.192	.949/.096	.891/.326	.934/.175	.945/.089
64	.954/.177	.948/.089	.950/.044	.888/.158	.936/.085	.946/.043
128	.949/.124	.949/.062	.949/.031	.893/.111	.938/.060	.947/.031
256	.951/.087	.952/.044	.951/.022	.899/.078	.939/.042	.950/.022
$\rho = .10$						
16	.948/.376	.948/.190	.949/.095	.886/.322	.934/.173	.947/.088
64	.952/.175	.950/.088	.950/.044	.898/.157	.935/.085	.946/.043
128	.953/.123	.949/.061	.946/.031	.893/.110	.938/.060	.943/.030
256	.949/.086	.951/.043	.950/.022	.892/.077	.940/.042	.947/.021
$\rho = .30$						
16	.952/.345	.943/.173	.932/.087	.898/.298	.935/.161	.932/.082
64	.952/.161	.946/.081	.945/.040	.894/.145	.936/.078	.942/.040
128	.949/.113	.949/.056	.946/.028	.894/.101	.937/.055	.944/.028
256	.950/.079	.950/.040	.948/.020	.889/.071	.935/.039	.943/.020
$\rho = .50$						
16	.948/.283	.935/.142	.885/.071	.890/.252	.930/.136	.894/.070
64	.948/.133	.950/.066	.940/.033	.890/.119	.938/.065	.934/.033
128	.949/.093	.950/.046	.947/.023	.891/.084	.938/.045	.945/.023
256	.950/.065	.951/.033	.949/.016	.889/.059	.937/.032	.946/.016
$\rho = .70$						
16	.944/.190	.920/.095	.828/.047	.891/.176	.930/.097	.845/.050
64	.949/.090	.942/.045	.920/.022	.887/.082	.931/.044	.923/.023
128	.951/.063	.946/.032	.934/.016	.893/.057	.934/.031	.937/.016
256	.951/.044	.947/.022	.942/.011	.895/.040	.935/.022	.944/.011
$\rho = .90$						
16	.940/.070	.898/.035	.741/.017	.885/.068	.920/.038	.786/.020
64	.949/.033	.938/.017	.909/.008	.894/.031	.930/.017	.912/.009
128	.951/.024	.945/.012	.929/.006	.892/.021	.935/.012	.925/.006
256	.950/.017	.945/.008	.942/.004	.887/.015	.935/.008	.938/.004

Note. Numbers on the left of the slashes indicate coverage rates, numbers on the right are estimates of interval widths. N = Sample size per study, HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach.

comes at the cost of coverage rates being smaller than desired. Second, for $k = 8$ the HS approach always shows coverage rates less than .90. In these situations, the differences in interval widths to the HO/RR approach are also at a maximum. The inferior performance of the HS approach may be due to second order sampling error in the estimated standard errors for the mean effect sizes when the number of studies is small.

It is also possible on the basis of the results shown in Table 2.3 to shed some light on the performance of the approaches for significance testing, as is ordinarily done in meta-analyses. From the results on the estimation of the MES in Table 2.2, it is known that biases are generally small, so that interval widths can be used to deduce the following results: First, because of smaller interval widths for the HS approach, the hypothesis in question will be rejected more often as in the HO/RR approach. Thus, statistical power will be comparatively higher for the HS approach when the hypothesis is false but the HO/RR may better conform to the desired nominal α-level when the hypothesis is true. Second, for small effects ($\rho = .10$) and combinations of small to medium levels of k and N there is remarkably low power for both approaches. This can be seen, for example, by considering the situation in which $k = 8$, $N = 64$, and $\rho = .10$. Here, there is only a very small bias for both approaches (.001 in absolute value, see Table 2.2) and interval widths are .175 for the HO/RR approach and .157 for the HS approach, respectively. Thus, intervals will mostly contain the value of zero so that the hypothesis is (falsely) not rejected. The results we deduced on testing the hypothesis of zero correlation in the population, conform to the conclusions drawn by Field (2001), who explicitly examined the test performance of the approaches but not the performance for confidence intervals as we have done.

In sum, the HO/RR approach shows a better overall performance for the confidence intervals in comparison to the HS approach. The most disturbing aspect of the results for the HS approach is the consistently low coverage rate for a small number of studies in a meta-analysis. Although most meta-analyses in practice will have more than eight studies in total, this result is particularly relevant for subgroup analyses often done in detailed analyses of a meta-analytic database.

The last aspect in the comparative evaluation of the approaches is the performance of the proposed homogeneity tests. The estimates of the Type I error rates for homogeneity tests are shown in Table 2.4. The results indicate that the test in the HO/RR approach approximately retains the nominal α (.05) in all situations. In contrast, the results for the HS approach show deficiencies as the population correlation increases, N is small, and k grows larger. This is most easily noticeable by inspecting the rejection rates of the null hypothesis for $\rho = .90$. Again, it is noted that this situation will not be given often in practice. Nevertheless, the deviance of the results from the expected values for the HS approach and its conspicuously low rejection rates for small N and ρ lead to a preference for the HO/RR approach in the situations of the simulation study.

Table 2.4 Results: Type I Error Rates for Homogeneity Tests ($\alpha = .05$)

	Number of studies k					
	HO/RS			HS		
N	8	32	128	8	32	128
$\rho = 0$						
16	.057	.053	.054	.044	.035	.034
64	.050	.053	.053	.046	.048	.047
128	.052	.051	.050	.052	.049	.047
256	.053	.051	.051	.052	.050	.050
$\rho = .10$						
16	.056	.053	.053	.044	.037	.037
64	.050	.051	.048	.048	.047	.043
128	.051	.050	.053	.048	.049	.050
256	.049	.052	.049	.048	.052	.049
$\rho = .30$						
16	.051	.055	.054	.045	.049	.056
64	.049	.054	.055	.050	.053	.054
128	.049	.052	.049	.048	.049	.051
256	.048	.049	.055	.047	.049	.054
$\rho = .50$						
16	.053	.050	.048	.062	.079	.124
64	.047	.054	.052	.051	.062	.068
128	.048	.048	.051	.050	.051	.059
256	.050	.053	.050	.051	.054	.053
$\rho = .70$						
16	.049	.048	.040	.079	.138	.247
64	.049	.046	.048	.057	.064	.088
128	.048	.050	.048	.052	.064	.068
256	.054	.048	.048	.055	.055	.057
$\rho = .90$						
16	.048	.043	.034	.097	.210	.445
64	.053	.045	.046	.065	.081	.125
128	.050	.048	.051	.055	.066	.085
256	.048	.050	.050	.050	.061	.063

Note. N = Sample size per study, HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach.

2.4 DISCUSSION AND CONCLUSIONS

From the findings in our simulation study we conclude that the HO/RR approach leads to more reliable results in a meta-analysis of correlation coefficients in a homogeneous situation. Most of our predictions were supported by the results. For the accurate estimation of a mean effect size it does not seem to be critical which of the two approaches is chosen because estimates were mostly accurate within rounding error for both approaches. However, for the construction of confidence intervals or testing of the hypothesis that the population correlation is zero, the HO/RR approach leads to more appropriate results. The same is true for testing the hypothesis of a constant effect size in the population with the homogeneity test (but see Hartung et al., Chapter 1 in this volume).

There are two aspects of the simulation study that might limit its relevance for practical purposes. First, only the homogeneous case was investigated. It has been argued by several authors (e.g., Erez et al., 1996; Hunter & Schmidt, 2000; Osburn & Callender, 1992) that the assumption of homogeneous effect sizes is not realistic for most practical applications. However, despite several calls for an increased use of random effects procedures, methods of meta-analysis as presented in this chapter are still dominant in the literature in the social sciences. For an evaluation of the methods most often used in practice, it seems reasonable to compare them in simulations of situations for which the procedures were designed, as was done in the present study. Our results showed that in this case the HO/RR approach is preferable to the HS approach.

It might well be the case, however, that the procedures perform differently in heterogeneous situations. Whereas the HS approach is supposed to be applicable both in homogeneous and heterogeneous situations, there have been developed different procedures for the two situations by Hedges and Olkin (1985). That is, for the application of the procedures presented by Hedges and Olkin one has to make a decision between procedures *before* their application and this decision depends on assumptions about the true situation in the population of effect sizes. Field (2001) presented a simulation study in which the random effects procedures by Hedges and Olkin were compared to the HS approach in heterogeneous situations. He reported mixed results insofar as there are advantages in using the HS approach for estimation of the mean effect size but also a better performance of the random effects procedures by Hedges and Olkin for inferential purposes, though none of the approaches performed satisfyingly for all simulated conditions. In a comprehensive effort to compare a large set of approaches and accompanying refinements in a number of situations, Schulze (in press) analyzed the approaches as presented here also in heterogeneous situations. In this study, a series of serious deficiencies of both approaches in heterogeneous situations was found and better alternatives were pointed out.

As an alternative to the above mentioned a priori decision between fixed and random effects procedures, Hedges and Vevea (1998) also proposed a so-called conditionally random effects procedure in which the choice of procedures is conditional on the results of the homogeneity test. They showed analytically that their fixed and random effects procedures perform best with respect to inferential goals when applied to situations for which they were designed. The conditionally random effects procedure showed performance better than the random effects procedure in homogeneous situations but performed not as good as the fixed effects procedures here. The pattern of performance was reversed for heterogeneous situations. Several simulation studies (e.g., Field, 2001; Hardy & Thompson, 1998; Harwell, 1997; Schulze, in press) showed, however, that there are serious problems with the homogeneity test and that it should be used with caution as a device to decide between models.

In sum, at least for inferential purposes the decision between fixed and random effects procedures is critical because they perform best when the decision is correct. Thus, such a decision is important when a comparison of approaches is of interest. Unfortunately, authoritative statistical tests for an empirically based decision do not yet seem to be available and compromise procedures like the conditional random effects model or the HS-approach are not without problems. Additionally, a choice between models also crucially depends on the inferential purposes as Hedges and Vevea (1998) have argued and we share their view that an abandonment of fixed effects procedures – even if heterogeneous situations are assumed by default – would be unnecessary.

A second potential limitation for the conclusions based on the results presented in this chapter is that the sample sizes were held constant in the simulations within levels of k and ρ. Indeed, constant sample sizes have never been observed in published meta-analyses and are therefore not realistic. We argue, however, that if differences in sample sizes would exhibit an influence on the results of a meta-analysis, such an influence would depend on the distribution of the sample sizes, its parameters, and also the covariation of sample sizes with effect sizes. Furthermore, this influence might differentially affect the results for the approaches to be compared. If this would be the case, specific assumptions about the distributional properties of sample sizes would obscure a comparative evaluation of approaches. Only in a situation in which firm knowledge about the distribution and its properties were available there would be a gain in making the comparison of approaches more realistic. However, there seems to be no consensus about which distribution to assume in simulation studies. In some Monte-Carlo studies on meta-analytical methods, the sample sizes across studies have been assumed to be normally distributed with varied parameters (e.g., Field, 2001; Osburn & Callender, 1992), uniformly distributed (e. g., Erez et al., 1996), or have been held constant (e. g., Corey et al., 1998; Overton, 1998) as in the present study. Unfortunately, neither of these distributions seems to mirror the distribution observed in practice. The results of a content analysis of 81 meta-analyses in industrial/organizational psychology reported by Cornwell (1988), clearly show a distribution of sample sizes far from normal or uniform. Instead, at least in this field of research the distribution is characterized by strong positive skewness and kurtosis. We therefore think that holding sample sizes constant across studies is a more sensible choice over assuming a distribution not observed in practice.

Notwithstanding the outlined potential limitations, it is clear from the results presented in this chapter that many of the conclusions concerning the HS approach drawn by Johnson et al. (1995) were based on erroneous results. However, more extensive simulation studies are needed to reach final conclusions about the usefulness of existing approaches.

REFERENCES

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, *99*, 388–399.

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*, 825–840.

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations. *The Journal of General Psychology*, *125*, 245–261.

Cornwell, J. M. (1988, August). *Content analysis of meta-analytic studies from I/O psychology.* Paper presented at the 96th annual meeting of the American Psychological Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED 304469).

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, *49*, 275–306.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*, 71–79.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte-Carlo comparison of fixed- and random effects-methods. *Psychological Methods*, *6*, 161–180.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

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

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods*, *2*, 219–231.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* London: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, *15*, 193–232.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*, 275–292.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, *80*, 94–106.

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 3–10). New York: Russell Sage Foundation.

Osburn, H. G., & Callender, J. C. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, *77*, 115–122.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*, 354–379.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, *86*, 1165–1168.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

Schmidt, F. L., & Hunter, J. E. (1995). The impact of data-analysis methods on cumulative research knowledge. *Evaluation & The Health Professions*, *18*, 408–427.

Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology*, *84*, 144–148.

Schulze, R. (in press). *Meta-analysis: A comparison of approaches.* Seattle, WA: Hogrefe & Huber.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, *72*, 146–148.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research.* Chichester: Wiley.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

3

Meta-Analysis: A General Principle for Estimating Heterogeneity Variance in Several Models

Uwe Malzahn

Working Group: Biometry and Epidemiology
Institute for International Health, Joint Center for Humanities and Health Sciences
Free University Berlin

Summary

A main question in meta-analysis is the comparability of studies in consideration. This relates and leads inevitably to the investigation of problems of heterogeneity. In this chapter, we deal with the one-dimensional case, represented by four examples, and propose a nonparametric moment estimator for the heterogeneity variance in the corresponding random effects model. The principle is based on decomposing the variance of the study estimator, that is, the total (unconditional) variance is composed of the mean conditional variance and the heterogeneity variance, or an expression containing the latter. We also hint to problems concerning the use of the DerSimonian-Laird estimator, which is a frequently used nonparametric estimator of general application. Finally, based on expressions for the conditional variances in several models for effect parameters and quality scores we demonstrate our principle.

3.1 INTRODUCTION AND EXAMPLES

We denote by $\theta \in \mathbb{R}$ the measure of the effect of interest. Given population heterogeneity, there is an increase of the variance for the study estimator. It appears an additional variance term, the heterogeneity variance τ^2. An estimation $\hat{\tau}^2$ for τ^2 can be used to adapt the inference regarding the overall mean of θ.

We suppose that the meta-analysis is based on k studies, or charges of a pharmaceutical product (solution, powder), respectively. Let us consider four examples for an effect measure or quality score θ.

Standardized difference This effect size measure is used for comparisons of groups based on continuous measurement variables with (possibly) different scales of measurement:

$$\theta = \frac{(\mu^T - \mu^C)}{\sigma_{T,C}}.$$

Here, μ^T and μ^C denote the mean values in a treatment and control group, $\sigma_{T,C}$ denotes the variance of the response variables in the two groups. That means equal variances of the two groups within each study are assumed.

Standardized mortality ratio Here, we are interested in the expected number of counts for a case event in a region or area in comparison to the corresponding number in a reference population with the same population structure:

$$\theta = \frac{\mu}{e}.$$

At this μ is the mean number of mortality or morbidity cases for a geographic region or area, and e is the corresponding value for this area calculated on the basis of an external reference population. Clearly, both values depend on the population size for the area considered.

Log relative risk in Cox regression with random censorship We consider the log-linear Cox-model with only one covariate. This covariate is a dichotomous variable which indicates some grouping membership (new therapy/treatment – standard therapy/placebo). Here, the parameter θ can be interpreted as the *logarithm of the relative risk* ($\ln(RR)$):

$$\theta = \ln(RR) = \ln\left(\frac{\lambda(t|Z=1)}{\lambda(t|Z=0)}\right) = \ln\left(\frac{\lambda^T(t)}{\lambda^C(t)}\right),$$

where $\lambda(\cdot|Z) = \lambda_0(\cdot)\mathrm{e}^{\theta Z}$ denotes the hazard rate function. We suppose that the distribution function $F_0(t) = P(T \leq t|Z=0)$ is continuous.

$S_0(t) = \exp\left(-\int_0^t \lambda_0(z)dz\right)$ is the survival function corresponding to $\lambda_0(\cdot)$.

Quality scores (in pharmaceutical technology) This type of scores is used for in-process control detecting polluting particles in solutions and powders:

$$\theta = \frac{1}{n} c_0 \left(\lambda_l + c_m \lambda_m + c_g \lambda_g \right).$$

Here, n denotes the size of the sample which is drawn from each charge of the product; λ_l, λ_m, and λ_g denote the expected numbers of slight, moderate, and severe faults (pollutions, contaminations) in the sample; c_m and c_g are coefficients to weigh the kinds of fault.

The (unknown) study-/charge-specific values of the effect-/score parameter are denoted by θ_i; $\hat{\theta}_i$ is the estimate in the study number i. Homogeneity means that $\theta_1 = \theta_2 = \cdots = \theta_k$.

In the random effects model (RE model) we have to distinguish between the conditional distribution of the random variable $\hat{\theta}$, given a fixed study-specific parameter value θ, $P^{\hat{\theta}}$, and the distribution of the parameter θ in the population of study parameters. In the context of heterogeneity analysis the latter distribution is called the heterogeneity distribution, say G.

We will assume that $\hat{\theta}$ is a conditional unbiased estimator[1], and $\hat{\theta}$ is regarded as the bias corrected version of an estimator $\tilde{\theta}$. In the following, we give the study estimators for our four examples.

Standardized difference

$$\hat{\theta}_i = \left[H(N_i/2)\right]^{-1} \frac{\left(\overline{X}_i^T - \overline{X}_i^C\right)}{s_i^2}.$$

Here, s_i^2 denotes the pooled sample variance, and $H(N_i/2)$ is the bias correcting factor:

$$H(a) = \sqrt{a}\frac{\Gamma(a-\frac{1}{2})}{\Gamma(a)}, \text{ and furthermore, } N_i = n_i^T + n_i^C - 2,$$

in which n_i^T and n_i^C are the group sizes in study i, and $\Gamma(\cdot)$ denotes the gamma function.

Standardized mortality ratio

$$\hat{\theta}_i = \frac{Y_i}{e_i}.$$

Here, Y_i is the observed number of mortality or morbidity cases in region i, and e_i is the corresponding expected number calculated on the basis of an external reference population.

[1] More precisely, we only need that $E\left(\hat{\theta}|\theta\right) = \theta + C$, in which the constant C does not depend on θ.

Log relative risk in Cox regression with random censorship Here, $\hat{\theta}_i$ is the maximum partial likelihood estimator (MPLE), which is asymptotically unbiased:

$$\hat{\theta}_i = \text{argmax}_\theta \prod_{j=1}^{L_i} \frac{\exp(\theta_i Z_{i(j)})}{\sum_{l \in R_{ij}} \exp\left(\theta_i Z_{il}\right)}$$

and n_i denotes the number of individuals at the beginning of study i, $T_{i1}^0 < \cdots < T_{iL_i}^0$ are the failure times, and $Z_{i(j)}$ is the covariate value for the individual failed at time T_{ij}^0. Furthermore, R_{ij} is the risk set immediately before time T_{ij}^0 in study i. The data are (X_{il}, δ_{il}), $X_{il} = \min(T_{il}, U_{il})$, in which T_{il} and U_{il} are the variables for the failure time and censoring time for individual l in study i, and $\delta_{il} = I_{\{T_{il} \leq U_{il}\}}$.

Quality scores

$$\hat{\theta}_i = \frac{c_0}{n_i}\left(l_i + c_m m_i + c_g g_i\right),$$

where l_i, m_i and g_i are the observed numbers of slight, moderate, and severe faults in charge i. Here, we usually have $n_i \equiv n$.

Let us summarize the assumptions of the RE model:

$$\hat{\theta}_i = \theta_i + \varepsilon_i, \quad \theta_i = \mu_G + \xi_i, \tag{3.1}$$

in which the ε_i are independent random variables with

$$\begin{aligned} E\left(\varepsilon_i\right) = 0, \quad \nu_i^2 = &\quad \text{Var}\left(\varepsilon_i\right) = E_G\left(\sigma_i^2\left(\theta_i\right)\right), \text{with } \sigma_i^2\left(\theta_i\right) = \text{Var}\left(\hat{\theta}_i|\theta_i\right), \\ \xi_i \text{ i.i.d., } E\left(\xi_i\right) = 0, &\quad \text{Var}\left(\xi_i\right) = \text{Var}_G\left(\xi\right) = \tau^2, \quad \mu_G = E_G\left(\theta_i\right). \end{aligned}$$

The conditional variances possibly depend both on the study design in study i and on the parameter θ_i. We consider the problem of estimating the heterogeneity variance τ^2.

3.2 A VARIANCE DECOMPOSITION

In this section, we generally denote by $\hat{\theta}$ a study-/charge-estimator with

$$\begin{aligned} \mu(\theta) &= E\left(\hat{\theta}|\theta\right), \\ \sigma^2(\theta) &= \text{Var}\left(\hat{\theta}|\theta\right). \end{aligned}$$

Note, that more precisely we have to denote $\sigma^2(\theta; \Xi; \alpha_1, \cdots, \alpha_p)$, in which Ξ stands for the study design or, more generally, for characteristics of the experiment, for instance $N = n^T + n^C - 2$ for the standardized difference. Under the assumption θ random, $\theta \sim G$, we can decompose the total (unconditional) variance:

$$\begin{aligned}\operatorname{Var}\left(\hat{\theta}\right) &= E_G\left(\operatorname{Var}\left(\hat{\theta}|\theta\right)\right) + \operatorname{Var}_G\left(E\left(\hat{\theta}|\theta\right)\right) \\ &= \int \sigma^2\left(\theta\right) g\left(\theta\right) d\theta + \int \left(\mu\left(\theta\right) - \mu_G\right)^2 g\left(\theta\right) d\theta,\end{aligned} \tag{3.2}$$

where $g(\cdot)$ is the density or the probability mass function (which gives the single probabilities) in the case of a discrete heterogeneity distribution.

In the case that $\hat{\theta}$ is conditionally unbiased: $\mu(\theta) = \theta$, or if $\mu(\theta) = \theta + \text{const.}$, it follows that $\operatorname{Var}_G\left(E\left(\hat{\theta}|\theta\right)\right) = \operatorname{Var}_G\left(\theta\right) = \tau^2$, and

$$\tau^2 = \operatorname{Var}\left(\hat{\theta}\right) - E_G\left(\operatorname{Var}\left(\hat{\theta}|\theta\right)\right) \tag{3.3}$$

(see Equation 3.2). Equation 3.3 will motivate a principle for estimating τ^2. The advantages of the method are:

- the resulting estimator is very easily calculated,
- we avoid any parametric assumption about G,
- using $\operatorname{Var}\left(\hat{\theta}|\theta\right)$, it is possible to take the special statistical model into account, that is, the special estimating problem.

This method is applicable under the supposition that we can express

$$E_G\left(\operatorname{Var}\left(\hat{\theta}|\theta\right)\right) = F\left(\Lambda; \mu_G^{(l)}; E_G\left(\alpha_s^r\right)\right), \tag{3.4}$$

in which Λ comprises known quantities from the study design. Examples for α_s are $\alpha_1 = p_l$, $\alpha_2 = p_m$, and $\alpha_3 = p_g$ in the case of quality scores. $\sigma^2(\theta) = \tilde{F}(\Lambda; \theta^l; \alpha_s^r)$ is sufficient for Equation 3.4.

3.3 THE DERSIMONIAN-LAIRD ESTIMATOR

A simple, general, and frequently used method to estimate τ^2 is the *DerSimonian-Laird estimator*. Generally, this estimator can be derived without normality assumptions by means of the weighted least squares principle. For this, it is assumed that the conditional variances are *known*, the so-called study specific variances. We write ν_i^2 instead of $\sigma_i^2(\theta_i)$ because it makes no sense to assume on the one side that θ_i is an unknown realization of a random variable, furthermore σ_i^2 is known exactly, and on the other side, that we have a structural dependence of σ_i^2 on $\theta_i : \sigma_i^2(\theta_i)$.

Now we can write the model (see Equations 3.1) in vector notation with the "design matrix" $X = \mathbf{1}_k = (1, \ldots, 1)^T$:

$$\left(\hat{\theta}_1, \ldots, \hat{\theta}_k\right)^T = D = X\beta + F = \mu_G \mathbf{1}_k + (E + C),$$

in which

$$\begin{aligned} E &= (\varepsilon_1, \ldots, \varepsilon_k)^T, \\ C &= (\xi_1, \ldots, \xi_k)^T, \\ E(E+C) &= \mathbf{0}_k, \\ W := \operatorname{Cov}(E+C) &= \tau^2 \mathbf{I}_k + V. \end{aligned}$$

Here, $\mathbf{I}_k$ denotes the k-dimensional identity matrix and $V = \operatorname{diag}(\nu_1^2, \ldots, \nu_k^2)$. Then it is straightforward to derive the weighted least squares (WLS) estimate $X\hat{\beta}_{\mathrm{WLS}} = \hat{\mu}_G$ with weighting matrix V^{-1} (note, that τ^2 is *unknown* but the ν_i^2 are assumed to be *known*). We have

$$\hat{\mu}_G = \left(\sum_{i=1}^{k} \nu_i^{-2} \hat{\theta}_i \right) / \left(\sum_{i=1}^{k} \nu_i^{-2} \right).$$

The corresponding sum of squared residuals is

$$RSS = |D - \hat{D}|^2_{V^{-1}} = \sum_{i=1}^{k} \nu_i^{-2} \hat{\theta}_i^2 - \left(\left(\sum_{i=1}^{k} \nu_i^{-2} \hat{\theta}_i \right)^2 / \left(\sum_{i=1}^{k} \nu_i^{-2} \right) \right),$$

with

$$E(RSS) = (k-1) + \left(\sum_{i=1}^{k} \nu_i^{-2} - \left(\sum_{i=1}^{k} \nu_i^{-4} / \sum_{i=1}^{k} \nu_i^{-2} \right) \right) \tau^2.$$

Rearranging this equation and replacing the expectation $E(RSS)$ by its observed value RSS, it follows

$$\hat{\tau}_{dl}^2 = \frac{(RSS - (k-1))}{(S_1 - (S_2/S_1))} = \frac{\left(\sum_{i=1}^{k} \nu_i^{-2} \left(\hat{\theta}_i - \hat{\mu}_G \right)^2 - (k-1) \right)}{(S_1 - (S_2/S_1))}, \tag{3.5}$$

with $S_l = \sum_{i=1}^{k} \left(\nu_i^{-2} \right)^l$, $l = 1, 2$.

However, for the application in practice the true study specific variances ν_i^2 are *unknown*, that means, for the data analysis they are *estimated*. Additionally, in most applications we have $\operatorname{Var}\left(\hat{\theta}|\theta\right) = \sigma^2(\theta)$ and $\sigma^2\left(\alpha_1, \ldots, \alpha_p\right)$, respectively, with unknown θ and α_s. Note, that in Equation 3.5 we have $\hat{\mu}_G = \hat{\mu}_G(\nu_1^2, \ldots, \nu_k^2)$ and $S_l = S_l(\nu_1^2, \ldots, \nu_k^2)$.

The problem is: What do we have to put in for $(\nu_1^2, \ldots, \nu_k^2)$? It would be good practice to start from an adequate model, find the right conditional distribution in this model, and with this derive the expression for the conditional variances $\sigma_i^2(\theta_i)$. Finally, we put in

$$\nu_i^2 := \hat{\sigma}_i^2(\theta_i) = \sigma_i^2\left(\hat{\theta}_i\right),$$

to obtain a practical version of the DerSimonian-Laird estimator. Note, that $\hat{\theta}_i$ is random. Consequently, $\hat{\tau}^2_{dl}$ is no longer the best linear unbiased estimator because the optimal weights ν_i^{-2} are unknown. Moreover, τ^2_{dl} is not unbiased, numerator *and denominator* in Equation 3.5 are stochastic terms.

3.4 THE CONDITIONAL VARIANCES IN THE MODELS

For the Examples

Standardized difference Under the assumption of normally distributed measurement variables

$$X_{ij} \sim \mathcal{N}(\mu_i^T, \sigma^2_{i;T,C}),\ j = 1, \ldots, n_i^T,$$
$$Y_{ij} \sim \mathcal{N}(\mu_i^C, \sigma^2_{i;T,C}),\ j = 1, \ldots, n_i^C$$

it follows that

$$\sqrt{q_i} H\left(N_i/2\right) \hat{\theta}_i \sim t_{N_i}\left(\theta_i \sqrt{q_i}\right),$$

a noncentral t-distribution with N_i degrees of freedom and noncentrality parameter $\theta_i \sqrt{q_i}$, in which $q_i = n_i^T n_i^C / \left(n_i^T + n_i^C\right)$. Therefore, we have

$$\sigma_i^2\left(\theta_i\right) = \left(H\left(N_i/2\right)\right)^{-2} \frac{N_i}{q_i\left(N_i - 2\right)} + \left(\frac{N_i}{\left(N_i - 2\right)}\left(H\left(H_i/2\right)\right)^{-2} - 1\right)\theta_i^2 \quad (3.6)$$

(see Malzahn, Böhning, & Holling, 2000).

Standardized mortality ratio Since conditional on the value θ_i in area i, a Poisson distribution with parameter $\lambda_i = \theta_i e_i$ is assumed for Y_i, it is easy to see that

$$\sigma_i^2 = \frac{\theta_i}{e_i}.$$

Maximum partial likelihood estimator for survival time studies It can be shown (see Fleming & Harrington, 1994) that

$$n_i^{1/2}\left(\hat{\theta}_i^{(n_i)} - \theta_i\right) \longrightarrow_L X,$$

with $X \sim \mathcal{N}\left(0, \sigma_i^{-2}\left(\theta_i\right)\right)$, in which the inverse of the asymptotical variance for the standardized estimator is under additional assumptions (in order to reduce the complexity of the resulting expression):

$$\begin{aligned}\sigma_i^2(\theta_i) &= \frac{1}{2}\exp(\theta_i)\int_0^{t_{up}}\left(1-\frac{t}{u_i}\right)\frac{(S_0(t))^{\exp(\theta_i)}}{\left[\exp(\theta_i)(S_0(t))^{\exp(\theta_i)}+S_0(t)\right]}f_0(t)\,dt\\ &= \frac{1}{2}\exp(\theta_i)\int_{S_0(t_{up})}^{1}\left(1-\frac{S_0^{-1}(s)}{u_i}\right)\frac{s^{\exp(\theta_i)}}{\left[\exp(\theta_i)s^{\exp(\theta_i)}+s\right]}ds.\end{aligned}\tag{3.7}$$

Here, t_{up} denotes a constant, common for all studies (it depends on the baseline hazard rate function, which is taken as a basis), and u_i are (possibly study-specific) constants, characterizing the censoring time distribution; S_0 denotes the survival function according to the baseline hazard: $S_0(s) = P_0(T > s)$.

Quality scores The situation can be described by a multinomial distribution model $M(n; p_l, p_m, p_g)$ with probability mass function:

$$\pi(l,m,g|\mathbf{p}) = \frac{n!}{l!m!g!}p_l^l p_m^m p_g^g(1-p_l-p_m-p_g)^{n-l-m-g}. \tag{3.8}$$

Here, $\mathbf{p} = (p_l, p_m, p_g)^T$ denotes the vector of the probabilities for detecting a slight, moderate, or severe contamination in an inspected item. For heterogeneity analysis it is important that we interpret the expression in Equation 3.8 as a *conditional* distribution: the distribution for the vector $(l,m,g)^T$ at fixed underlying vector $(p_l, p_m, p_g)^T$. Heterogeneity means: There exists a non-degenerated heterogeneity distribution G on $(0,1)^3$, and for each charge of the product under examination the actually underlying parameter vector $\mathbf{p}$ is a realization from this distribution. In this model, the conditional[2] variance of the quality score in charge i is

$$\begin{aligned}\sigma^2\left(\mathbf{p}^{(i)}\right) &:= \mathrm{Var}\left(\hat{\theta}_i|\mathbf{p}^{(i)}\right)\\ &= \frac{c_0^2}{n}\left[p_l^{(i)}\left(1-p_l^{(i)}\right)+c_m^2 p_m^{(i)}\left(1-p_m^{(i)}\right)+c_g^2 p_g^{(i)}\left(1-p_g^{(i)}\right)\right.\\ &\qquad\left.-2c_m p_l^{(i)}p_m^{(i)}-2c_g p_l^{(i)}p_g^{(i)}-2c_m c_g p_m^{(i)}p_g^{(i)}\right].\end{aligned}$$

3.5 A PRINCIPLE TO ESTIMATE THE HETEROGENEITY VARIANCE

Our starting point here is the relationship given in Equation 3.3 for the heterogeneity variance. If θ^2 enters in the expression for $\mathrm{Var}\left(\hat{\theta}|\theta\right)$, then τ^2 enters in

[2] Conditional on fixed $\mathbf{p}^{(i)} = \left(p_l^{(i)}, p_m^{(i)}, p_g^{(i)}\right)^T$.

an analogous manner in the right side of Equation 3.3, leading to an equation for τ^2 which is specifically considered for the model. This equation provides the possibility to construct an estimator $\hat{\tau}^2$. As an example, we want to demonstrate this principle for the standardized difference (see Malzahn et al., 2000).

Standardized difference Here, Equation 3.6 together with the relationship $E_G\left(\theta^2\right) = \tau^2 + \mu_G^2$ yields

$$\int \sigma^2(\theta)\, g(\theta)\, d\theta = (H(N/2))^{-2} \frac{N}{q(N-2)} + \left[\frac{N}{(N-2)} (H(N/2))^{-2} - 1\right] \left(\tau^2 + \mu_G^2\right).$$

Applying this and rearranging leads to

$$\tau^2 = (H(N/2))^2 \frac{(N-2)}{N} \operatorname{Var}\left(\hat{\theta}\right) - q^{-1} - \left[1 - \frac{(N-2)}{N} (H(N/2))^2\right] \mu_G, \tag{3.9}$$

where $N = n^T + n^C - 2$ and $q = n^T n^C / (n^T + n^C)$.

The data are $\left(\hat{\theta}_1; N_1, q_1\right), \ldots, \left(\hat{\theta}_k; N_k, q_k\right)$. Equation 3.9 will motivate a nonparametric estimator $\hat{\tau}^2$. To estimate the first term of Equation 3.9, it seems to be reasonable to use a modified version of the usual empirical variance of the study estimators, considering the different degrees of freedom (N_i). We can estimate the mean value of the effect parameter in the overall population by

$$\hat{\mu}_{\hat{\theta}} = k^{-1} \sum_{i=1}^{k} \hat{\theta}_i$$

or, given estimates $\hat{v}_i^2$ of the study-specific variances for $\hat{\theta}_i$, the pooled estimator

$$\hat{\mu}_{\hat{\theta}} = \frac{\sum_{i=1}^{k} \hat{v}_i^{-2} \hat{\theta}_i}{\sum_{i=1}^{k} \hat{v}_i^{-2}}.$$

In the fixed effects model for known study-specific variances, the pooled mean is the best unbiased linear estimator for the first moment and should be used if the data indicates at most a small heterogeneity variance. In the case of large heterogeneity, the arithmetic mean should be preferred because the "true" weights within the pooled estimator are poorly estimated by noniterative procedures. Finally, in Equation 3.9 we estimate

$$\left[1 - \frac{(N-2)}{N} (H(N/2))^2\right] \mu_G^2$$

by the mean value of the corresponding study specific realizations. This leads to a nonparametric estimator of the heterogeneity variance given by

$$\hat{\tau}^2 = \frac{1}{(k-1)} \sum_{i=1}^{k} (1 - K_i) \left(\hat{\theta}_i - \hat{\mu}_{\hat{\theta}}\right)^2 - \frac{1}{k} \sum_{i=1}^{k} \frac{1}{q_i} - \frac{1}{k} \sum_{i=1}^{k} K_i \hat{\theta}_i^2,$$

where

$$K_i = 1 - (H(N_i/2))^2 \frac{(N_i - 2)}{N_i}.$$

The same mode of procedure yields corresponding estimators for the heterogeneity variance in the case of the standardized mortality ratio.

Standardized mortality ratio

$$\hat{\tau}^2 = \frac{1}{(k-1)} \sum_{i=1}^{k} \left(\hat{\theta}_i - \hat{\mu}_{\hat{\theta}}\right)^2 - \frac{1}{k} \hat{\mu}_{\hat{\theta}} \sum_{i=1}^{k} \frac{1}{e_i},$$

where $\hat{\mu}_{\hat{\theta}}$ is an estimator for the mean value of the parameter in the whole population. Typically, two estimators are considered:

the arithmetic mean

$$\hat{\mu}_{\hat{\theta}}^{(ar)} = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}_i$$

and the pooled mean

$$\hat{\mu}_{\hat{\theta}}^{(pool)} = \frac{\sum_{i=1}^{k} \hat{\theta}_i e_i}{\sum_{i=1}^{k} e_i}$$

(see Böhning, Sarol, & Malzahn, 2000).

Log relative risk in the log-linear Cox model with random censorship At first, we consider the inverse of the asymptotical conditional variance of the standardized estimator in this model, given by Equation 3.7. This quantity has to be estimated for each study. The expression in Equation 3.7 for $\sigma_i^2(\theta_i)$ contains the unknown survival function $S_0(t)$ corresponding to the baseline hazard. An obvious nonparametric estimator $\hat{S}_0(t)$ in study i is the Kaplan-Meier estimator

$$\hat{S}_0(t) = \prod_{j:T_{ij}^0 \le t} \frac{(\bar{Y}_{ij} - 1)}{\bar{Y}_{ij}},$$

in which

$$\bar{Y}_{ij} = \sum_{l=1}^{n_i} Y_l\left(T_{ij}^0\right), \quad \text{with} \quad Y_l(t) = I_{(X_l \ge t)}.$$

Furthermore, since there are no bindings, the Nelson estimator for the cumulative hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(z)dz$$

is given by

$$\hat{\Lambda}_0 = \sum_{j:T_{ij}^0 \le t} \frac{\delta_{ij}}{\overline{Y}_{ij}}.$$

Consequently, a natural estimator for an integral of the form $\int_a^b h_i(t)\lambda_0(t)dt$ is given by

$$\sum_{j:a<T_j^0<b} h_i\left(T_j^0\right)\frac{\delta_j}{\overline{Y}_j}.$$

Because of $f_0(s) = \lambda_0(s)S_0(s)$, we have

$$h_i(t) = \frac{\left(S_0\left(t\right)\right)^{\exp(\theta_i)+1}}{\left[\exp\left(\theta\right)\left(S_0(t)\right)^{\exp(\theta_i)} + S_0(t)\right]}\left(1 - \frac{1}{u_i}\right)$$

in Equation 3.7, leading to the estimator

$$\begin{aligned}\hat{\sigma}_i^2(\theta_i) =& \frac{\exp(\hat{\theta}_i)}{2}\sum_{j:T_{ij}^0 \le t_{up}} \frac{\left(\hat{S}_0\left(T_{ij}^0\right)\right)^{\exp(\hat{\theta}_i)+1}}{\left[\exp(\hat{\theta}_i)\left(\hat{S}_0\left(T_{ij}^0\right)\right)^{\exp(\hat{\theta}_i)} + \hat{S}_0\left(T_{ij}^0\right)\right]}\\ &\times\left(1 - \frac{T_{ij}^0}{u_i}\right)\frac{\delta_{ij}}{\overline{Y}_{ij}},\end{aligned}$$

where $\overline{Y}_{ij}$ is the size of the risk set at failure time T_{ij}^0, and $\delta_{ij} := I_{\left(T_{ij} \le U_{ij}\right)}$, that is, $\delta_{ij} = 1$ if the individual number j in study i is a failure, and $\delta_{ij} = 0$ if this individual is a censored observation. Because the MPLE $\hat{\theta}$ is asymptotically unbiased, Equation 3.3 suggests estimators of the form

$$\hat{\tau}^2 = \frac{1}{(k-1)}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\mu}_{\hat{\theta}}\right)^2 - \hat{E}_G\left(\left(n\hat{\sigma}^2\left(\hat{\theta}\right)\right)^{-1}\right),$$

where

$$\hat{E}_G\left(\left(n\hat{\sigma}^2\left(\hat{\theta}\right)\right)^{-1}\right) = H\left(n_1^{-1}\hat{\sigma}_1^{-2}\left(\hat{\theta}_1\right), \ldots, n_k^{-1}\hat{\sigma}_k^{-2}\left(\hat{\theta}_k\right)\right).$$

The most simple case is $H(x_1, \ldots, x_k) = \overline{x}$, but this does not make much sense, rather it seems to be more sensible to derive a weighted mean of the

$n_i^{-1}\hat{\sigma}_i^{-2}(\hat{\theta}_i)$. Currently, this is an open problem and will be subject of further research.

Quality scores Here we can derive:

$$\begin{aligned}
\hat{\tau}^2 = & \frac{1}{(k-1)} \sum_{i=1}^{k} \left(\hat{\theta}_i - \hat{\mu}_{\hat{\theta}}\right)^2 \\
& - \frac{c_0^2}{n(n-1)} \left[\overline{b} + c_m^2 \overline{m} + c_g^2 \overline{g}\right] + \frac{c_0^2}{n(n-1)} \left[\overline{l^2} + c_m^2 \overline{m^2} + c_g^2 \overline{g^2}\right] \\
& - 2\frac{c_0^2}{n^2} \left[c_m \frac{1}{(k-1)} \sum_{i=1}^{k} \left(l_i - \overline{l}\right) (m_i - \overline{m}) + c_g \frac{1}{(k-1)} \sum_{i=1}^{k} \left(l_i - \overline{l}\right) (g_i - \overline{g}) \right. \\
& \left. + c_m c_g \frac{1}{(k-1)} \sum_{i=1}^{k} (m_i - \overline{m}) (g_i - \overline{g}) \right],
\end{aligned}$$

where

$$\overline{l^\alpha} = \frac{1}{k} \sum_{i=1}^{k} l_i^\alpha, \quad \overline{l} = \overline{l^1}.$$

REFERENCES

Böhning, D., Sarol, J., & Malzahn, U. (2000). *Efficient non-iterative and nonparametric estimation of heterogeneity variance for the standardised mortality ratio.* (unpublished manuscript).

Fleming, T. R., & Harrington, D. P. (1994). *Counting processes and survival analysis.* New York: Wiley.

Malzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta–analysis. *Biometrika, 87,* 619–632.

4

An Alternative Test Procedure for Meta-Analysis

Joachim Hartung and Guido Knapp
Department of Statistics[†]
University of Dortmund

Summary

In this chapter, we show that a meta-analysis carried out in the random effects model is preferable to the fixed effects model. Especially in the normal mean case, as our simulation study indicates, the test of association in the FE model does not yield satisfactory results. If one prefers to use the commonly used methods, the choice between the FE and the RE model, which leads to the choice of the test statistic for the hypothesis of no association, is better based on the sign of the method of moments estimator of the between-study variance than on the test of homogeneity. But the use of the alternative test statistic originally proposed in (Hartung, 1999) is preferable concerning the significance level to all commonly used methods. The test is always carried out in the RE model, but it yields sufficiently good results if no heterogeneity is present. So, one does not have to choose between the FE and the RE model in advance. In the case of a small between-study variance, a combined test procedure involving the commonly used test in the FE model and Hartung's alternative test statistic may still improve the actual significance level of the test towards the prescribed one.

[†] Project "Meta-Analysis in Biometry and Epidemiology" (SFB 475) of the Deutsche Forschungsgemeinschaft (DFG).

4.1 INTRODUCTION

In this chapter, we focus our attention on the tests of association in the meta-analytic framework, that is, we want to judge if an overall treatment effect exists given the stochastically independent study-specific estimates of the treatment effect. This test is carried out either in the fixed effects model of meta-analysis assuming a homogeneous treatment effect over the studies or in the random effects model if heterogeneity of the study-specific treatment effects is present. Before applying the test of association one usually decides which of the two models one takes. Even in recent literature one can find the proposal that the choice of the model should be based on the test of homogeneity, cf. for instance Normand (1999). But the test of homogeneity in this context often has too low power to detect a deviation from the hypothesis of homogeneity and the false use of the fixed effects model, if heterogeneity is present, can lead to a dramatic increase of the Type I error rate of the commonly used test of association as pointed out for instance in Ziegler and Victor (1999). Moreover, the commonly used tests of association in the fixed effects model and in the random effects model, respectively, may lead to a large number of unjustified significant evidences even if one carries out the analysis in the correct model. In the fixed effects model this was shown by Li, Shi, and Roth (1994) and also Böckenhoff and Hartung (1998) in the normal mean case.

We will now consider an alternative test statistic for the test of association in the random effects model of meta-analysis proposed by Hartung (1999) and show that this test provides satisfactory results concerning the actual significance level in the fixed effects model as well as in the random effects model, so that with this test a choice between the two models, in advance, is unnecessary. Furthermore, we will discuss decision rules for the test of association, which combines the commonly used tests and this alternative test, and investigate these decision rules, whether they yield a further improvement with respect to the prescribed significance level.

The outline of the chapter is as follows: In the next two sections we first describe the theoretical foundations of the meta-analysis in a fixed effects and a random effects approach, respectively. In Section 4.4, the commonly used methods for a practical application in the fixed effects and random effects model are presented. Section 4.5 contains some results of the theoretical deficiency of the commonly used tests in the both models. In Section 4.6, the alternative test statistic in the random effects model proposed by Hartung (1999) is presented and in Section 4.7 some decision rules are discussed, which combine the commonly used tests and the alternative test. In a simulation study, of which the results are given in Section 4.8, the discussed tests are compared concerning their actual significance levels in the normal mean case. Finally, some conclusions are given.

4.2 THE HOMOGENEOUS FIXED EFFECTS MODEL

Let us consider k independent studies and let us denote by $\theta_1, \dots, \theta_k$ the (one-dimensional) parameters of interest, where each parameter stands for the treatment effect in a study. For example, the parameter θ_i, $i = 1, \dots, k$, may represent the mean and the standardized difference of means, respectively, for continuous outcome variables or the risk difference, the logarithmic odds ratio, and the relative risk, respectively, for binary outcome variables. In each study an estimate of the parameter θ_i, say $\hat{\theta}_i$, is available, and all study-specific estimators $\hat{\theta}_i$, $i = 1, \dots, k$, are stochastically independent. Assuming that the parameters of interest are fixed and homogeneous, that is, it holds $\theta_1 = \cdots = \theta_k = \theta$, and the study specific estimators $\hat{\theta}_i$ are at least approximately normally distributed and unbiased or at least consistent, then the so-called (homogeneous) fixed effects model (FE model) of meta-analysis is given by

$$\hat{\theta}_i \sim \mathcal{N}\left(\theta, \sigma^2(\hat{\theta}_i)\right), \quad i = 1, \dots, k, \tag{4.1}$$

where $\sigma^2(\hat{\theta}_i)$ denotes the variance of the estimator $\hat{\theta}_i$ in the ith study.

In model 4.1, the best linear unbiased estimator of the common mean θ is given by

$$\tilde{\theta}_{FE} = \sum_{i=1}^{k} \frac{v_i}{v} \hat{\theta}_i, \quad v = \sum_{i=1}^{k} v_i, \tag{4.2}$$

with $v_i = [\sigma^2(\hat{\theta}_i)]^{-1}$ the inverse of the variance of the study-specific estimator $\hat{\theta}_i$ in the ith study. The estimator $\tilde{\theta}_{FE}$ is also the maximum likelihood estimator (MLE) of θ in model 4.1 if the normal distribution exactly holds and the variances $\sigma^2(\hat{\theta}_i)$ are known.

The assumption of homogeneity of the parameters can formally be checked using the test statistic

$$Q = \sum_{i=1}^{k} v_i \left(\hat{\theta}_i - \tilde{\theta}_{FE}\right)^2, \tag{4.3}$$

which is at least approximately χ^2-distributed with $(k-1)$ degrees of freedom under the hypothesis of homogeneity (Cochran, 1954; Normand, 1999).

If all assumptions in model 4.1 are fulfilled, the estimator $\tilde{\theta}_{FE}$ from Equation 4.2 has the following distributional property:

$$\tilde{\theta}_{FE} \sim \mathcal{N}\left(\theta, \frac{1}{v}\right). \tag{4.4}$$

So, from 4.4 an (approximate) $(1-\alpha)$-confidence interval for the common parameter θ is given by $\tilde{\theta}_{FE} \mp u_{1-\alpha/2}/\sqrt{v}$, where u_γ denotes the γ-quantile of the standard normal distribution. Furthermore, the two-sided test rejects the hypothesis of no association $H_0 : \theta = 0$ at level α if

$$\frac{\left(\tilde{\theta}_{FE}\right)^2}{1/v} = \frac{\left(\sum_{i=1}^{k} v_i \hat{\theta}_i\right)^2}{v} > \chi^2_{1;1-\alpha},$$

where $\chi^2_{\nu;\gamma}$ denotes the γ-quantile of the χ^2-distribution with ν degrees of freedom.

If the assumption of homogeneity is not valid in model 4.1, that is, it exists at least one pair $\theta_i \neq \theta_j$, $i \neq j$, then the estimator $\tilde{\theta}_{FE}$ from Equation 4.2 is still an unbiased estimator of a weighted average of the θ_i's, namely of $\sum_{i=1}^{k} v_i \theta_i / v$. So, the above described confidence interval and test can always be used for this weighted average of the parameters. But the usual proceeding, if the hypothesis of homogeneity is not valid, is either to try to identify covariates which stratify studies into homogeneous populations or to carry out the meta-analysis in a random effects model (Normand, 1999). In the next section we will consider the latter proposal.

4.3 THE RANDOM EFFECTS MODEL

In contrast to the homogeneous fixed effects model 4.1, we first allow that the study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, may possess different expected values θ_i, $i = 1, \ldots, k$, that is, it holds approximately

$$\hat{\theta}_i | \theta_i, \quad \sigma^2(\hat{\theta}_i) \sim \mathcal{N}\left(\theta_i, \sigma^2(\hat{\theta}_i)\right), \quad i = 1, \ldots, k,$$

and for each study-specific mean θ_i we assume that it is drawn from some superpopulation of effects with mean θ and variance τ^2, that is,

$$\theta_i | \theta, \quad \tau^2 \sim \mathcal{N}\left(\theta, \tau^2\right).$$

The parameters θ and τ^2 are referred to as hyperparameters, θ represents the average treatment effect and τ^2 the between-study variation. Given the hyperparameters, the marginal distribution of the estimators $\hat{\theta}_i$ is given by

$$\hat{\theta}_i \sim \mathcal{N}\left(\theta, \tau^2 + \sigma^2(\hat{\theta}_i)\right), \quad i = 1, \ldots, k, \tag{4.5}$$

(cf. Whitehead & Whitehead, 1991; Normand, 1999). If the between-study variance τ^2 is equal to zero then the random effects model (RE model) 4.5 reduces to the FE model 4.1.

In the RE model 4.5, the best linear unbiased estimator of the average treatment effect θ is given by

$$\tilde{\theta}_{RE} = \sum_{i=1}^{k} \frac{w_i}{w} \hat{\theta}_i, \quad w = \sum_{i=1}^{k} w_i$$

with $w_i = [\tau^2 + \sigma^2(\hat{\theta}_i)]^{-1}$ the inverse of the variance of the estimator $\hat{\theta}_i$ in the RE model. The estimator $\tilde{\theta}_{RE}$ is also the MLE of θ if the variance components are known and the normal distribution in 4.5 exactly holds.

The estimator $\tilde{\theta}_{RE}$ in model 4.5 possesses the following distributional property:

$$\tilde{\theta}_{RE} \sim \mathcal{N}\left(\theta, \frac{1}{w}\right).$$

Thus, an $(1-\alpha)$-confidence interval for the average treatment effect θ in the RE model is given by $\tilde{\theta}_{RE} \mp u_{1-\alpha/2}/\sqrt{w}$ and the hypothesis of no association, that is, $H_0 : \theta = 0$, is rejected at level α if

$$\frac{\left(\tilde{\theta}_{RE}\right)^2}{1/w} = \frac{\left(\sum_{i=1}^{k} w_i \hat{\theta}_i\right)^2}{w} > \chi^2_{1;1-\alpha}.$$

4.4 THE COMMONLY USED METHODS IN THE FE AND RE MODEL

In the previous two sections we have summarized the theoretical aspects of the FE and RE model of meta-analysis. For a practical application of the just described inference the involved variances τ^2 and $\sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, are hardly ever known. So, they have to be replaced by appropriate estimators.

Let us first consider the FE model. Normally, an estimator of the variance of $\hat{\theta}_i$, say $\hat{\sigma}^2(\hat{\theta}_i)$, is available in each study. We assume that these estimators $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, are jointly stochastically independent and at least nearly unbiased for the corresponding variances $\sigma^2(\hat{\theta}_i)$. Using these variance estimators, the feasible estimator of θ in model 4.1 is given by

$$\hat{\theta}_{FE} = \sum_{i=1}^{k} \frac{\hat{v}_i}{\hat{v}} \hat{\theta}_i, \quad \hat{v} = \sum_{i=1}^{k} \hat{v}_i, \quad \hat{v}_i = \left[\hat{\sigma}^2(\hat{\theta}_i)\right]^{-1}, \quad i = 1 \ldots, k.$$

In general, this estimator is not an unbiased one of θ. If the estimators $\hat{\theta}_i$ and the variance estimators $\hat{\sigma}^2(\hat{\theta}_i)$ are stochastically independent, which implies that $\hat{v}_i/\hat{v}$ as a function of $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, is also stochastically independent of $\hat{\theta}_i$, then the estimator $\hat{\theta}_{FE}$ is unbiased for θ in model 4.1. This can be readily seen with $\sum_{i=1}^{k} \hat{v}_i/\hat{v} = 1$.

For further inference in the FE model, the variance estimators are commonly inserted in the test statistics and the confidence interval given in Section 4.2. Thus, the assumption of homogeneity of the treatment effects in model 4.1 is checked in practice with the test statistic

$$Q_1 = \sum_{i=1}^{k} \hat{v}_i \left(\hat{\theta}_i - \hat{\theta}_{FE}\right)^2. \tag{4.6}$$

The test statistic for testing the hypothesis of no association is given by

$$T_1 = \frac{\left(\hat{\theta}_{FE}\right)^2}{1/\hat{v}} = \frac{\left(\sum_{i=1}^{k} \hat{v}_i \hat{\theta}_i\right)^2}{\hat{v}}, \tag{4.7}$$

and the hypothesis is rejected at level α if the observed value of T_1 exceeds the $(1-\alpha)$-quantile of the χ^2-distribution with one degree of freedom.

In the RE model, besides the estimates of the within-study variances $\sigma^2(\hat{\theta}_i)$, $i = 1, \dots, k$, an estimator of the between-study variance τ^2 has to be deduced. One approach is to use the method of moments estimator, which is derived using the test statistic of homogeneity Q from Equation 4.3. The expected value of Q in the RE model is given by

$$E(Q) = (k-1) + \tau^2 \left(v - \sum_{i=1}^{k} v_i^2 / v \right)$$

(cf. DerSimonian & Laird, 1986; Whitehead & Whitehead, 1991). So, the method of moments estimator reads

$$\tilde{\tau}^2 = \frac{Q - (k-1)}{v - \sum_{i=1}^{k} v_i^2 / v}. \tag{4.8}$$

Due to its construction, the estimator $\tilde{\tau}^2$ is unbiased but it can yield negative estimates with positive probability. Moreover, the estimator depends on the unknown variances $\sigma^2(\hat{\theta}_i)$. Thus, for a practical application the feasible estimator of τ^2 is given by

$$\hat{\tau}^2 = \frac{Q_1 - (k-1)}{\hat{v} - \sum_{i=1}^{k} \hat{v}_i^2 / \hat{v}}, \tag{4.9}$$

with Q_1 from Equation 4.6, and usually the truncated version of this estimator is used, namely,

$$\hat{\tau}_+^2 = \max\{0, \hat{\tau}^2\}. \tag{4.10}$$

The larger the between-study variance τ^2, the less the probability of $\hat{\tau}^2$ yielding negative estimates. So, for large τ^2 both estimators in Equations 4.9 and 4.10 are nearly identical. Note that feasibility in this sense does not mean unbiasedness for all τ^2.

Other approaches of estimating the between-study variance τ^2 are to use the restricted maximum likelihood approach or a Bayesian approach (Normand, 1999). But we will not consider these approaches in the context of this chapter.

With the between-study variance estimator $\hat{\tau}_+^2$ and the within-study variance estimators $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \dots, k$, the feasible estimator of the average treatment effect θ in the RE model is given by

$$\hat{\theta}_{RE} = \sum_{i=1}^{k} \frac{\hat{w}_i}{\hat{w}} \hat{\theta}_i, \quad \hat{w} = \sum_{i=1}^{k} \hat{w}_i, \quad \hat{w}_i = \left[\hat{\tau}_+^2 + \hat{\sigma}^2(\hat{\theta}_i) \right]^{-1}, \; i = 1 \dots, k.$$

So, the test statistic for testing the hypothesis of no association in the RE model is given by

$$T_2 = \frac{\left(\hat{\theta}_{RE}\right)^2}{1/\hat{w}} = \frac{\left(\sum_{i=1}^{k} \hat{w}_i \hat{\theta}_i\right)^2}{\hat{w}}, \tag{4.11}$$

and the hypothesis is rejected at level α if the observed value of T_2 is larger than the $(1-\alpha)$-quantile of the χ^2-distribution with one degree of freedom.

4.5 THE THEORETICAL DEFICIENCY OF THE COMMONLY USED TESTS IN THE FE AND RE MODEL

In the previous section, we have presented the commonly used test statistics in the FE and RE model (see Equations 4.7 and 4.11) for testing the hypothesis of no association as well as the rule of rejection at a prescribed significance level α. But as already pointed out in Li et al. (1994) and Böckenhoff and Hartung (1998), the actual significance level of the commonly used test in the FE model with normally distributed responses is often much larger than the prescribed level α due to underestimation of the variance of the combined estimator $\tilde{\theta}_{FE}$ so that this phenomenon results in a large number of unjustified significant evidences.

We now summarize the main theoretical results of the work of Böckenhoff and Hartung in the FE model and indicate that the same deficiency also holds in the RE model. First, we need some mathematical tools from Hartung (1976).

Definition 4.5.1. A function $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is called convex if

$$\left(x, y \in \mathbb{R}^k, \lambda \in [0,1] \Rightarrow f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda) f(y)\right)$$

with the natural semi-ordering, that is,ordering by components.

Definition 4.5.2. A function $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is called quasi-convex if

$$\left(y \in \mathbb{R}^\ell \Rightarrow \{x \in \mathbb{R}^k | f(x) \leq y\} \text{ is convex}\right).$$

Definition 4.5.3. A function f is called (quasi-)concave if $(-f)$ is (quasi-)convex.

Lemma 4.5.1. *Let $f : \mathbb{R}^k \to \mathbb{R}^\ell$ be convex [concave] and $T : \mathbb{R}^\ell \to \mathbb{R}^m$ be (quasi-) convex [(quasi-)concave] and increasing by the natural semi-ordering in $\mathbb{R}^m$. Then the composed function $T \circ f$ is (quasi)-convex [(quasi)-concave].*

Lemma 4.5.2. *Let $f : \mathbb{R}^k \to \mathbb{R}^\ell$ be convex [concave] and $T : \mathbb{R}^\ell \to \mathbb{R}^m$ be (quasi-) concave [(quasi-)convex] and decreasing by the natural semi-ordering in $\mathbb{R}^m$. Then the composed function $T \circ f$ is (quasi)-concave [(quasi)-convex].*

Lemma 4.5.3. *If $f : \mathbb{R}_+^\ell \to \mathbb{R}_+$ is quasi-convex [quasi-concave] and $f(\lambda\, x) = \lambda\, f(x)$, $\lambda > 0, x \neq 0$, then f is also convex [concave].*

Jensen's Inequality. For a random variable $\hat{\vartheta}$ it holds
$E(f(\hat{\vartheta})) \geq f(E(\hat{\vartheta}))$ if f is convex, and
$E(g(\hat{\vartheta})) \leq g(E(\hat{\vartheta}))$ if g is concave.

We now return to the meta-analysis and consider the variance estimator $1/\hat{v}$ in the FE model. Note, that in the previous section we have assumed that the within-study estimators $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, are (nearly) unbiased for $\sigma^2(\hat{\theta}_i)$, that is, $\mathrm{E}(\hat{\sigma}^2(\hat{\theta}_i)) = \sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$. So, we can prove the following theorem.

Theorem 4.5.4. *In the FE model the variance estimator $1/\hat{v}$ in average underestimates the variance $1/v$, that is, $E(1/\hat{v}) \leq 1/v$.*

Proof. First consider $\hat{\sigma}^2(\hat{\theta}_i) > 0$, $i = 1, \ldots, k$, then $1/\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, is a convex function in $\hat{\sigma}^2(\hat{\theta}_i)$.

Furthermore, $\sum : \mathbb{R}^k \rightarrow \mathbb{R}$ is a convex and increasing function. So, with Lemma 4.5.1 the function $\sum_{i=1}^k 1/\hat{\sigma}^2(\hat{\theta}_i)$ is convex in $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$.

With Lemma 4.5.2, we yield that the function $\hat{v}^{-1} = \left(\sum_{i=1}^k 1/\hat{\sigma}^2(\hat{\theta}_i)\right)^{-1}$ is quasi-concave, because every monotone function is quasi-convex as well as quasi-concave.

Now, it holds for $\lambda > 0$ that

$$\left(\sum_{i=1}^k \frac{1}{\lambda\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1} = \left(\frac{1}{\lambda}\sum_{i=1}^k \frac{1}{\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1} = \lambda\left(\sum_{i=1}^k \frac{1}{\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1},$$

that means with Lemma 4.5.3 that the function $\left(\sum_{i=1}^k 1/\hat{\sigma}^2(\hat{\theta}_i)\right)^{-1}$ is concave. Applying Jensen's inequality, we obtain

$$E\left(\hat{v}^{-1}\right) = E\left(\sum_{i=1}^k \frac{1}{\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1} \leq \left(\sum_{i=1}^k \frac{1}{E(\hat{\sigma}^2(\hat{\theta}_i))}\right)^{-1} = \left(\sum_{i=1}^k \frac{1}{\sigma^2(\hat{\theta}_i)}\right)^{-1} = v^{-1},$$

which completes the proof. □

As a consequence of Theorem 4.5.4, we obtain that the distribution of the test statistic T_1 from Equation 4.7 under $H_0 : \theta = 0$ may not be well approximated by a χ^2-distribution with one degree of freedom. Suppose that the feasible estimator $\hat{\theta}_{FE}$ has variance near $1/v$ under H_0, then the probability of T_1 under H_0 to exceed the $(1 - \alpha)$-quantile of the χ_1^2-distribution is larger than α, so that the actual significance level of the test is larger than the prescribed one. Such an attitude of the test based on T_1 has been observed in the normal mean case as already mentioned by Li et al. (1994) as well as by Böckenhoff and Hartung (1998). In the latter work, some alternative estimators of the variance of $\tilde{\theta}_{FE}$ are discussed in the normal mean case, which in average overestimate the variance of $\tilde{\theta}_{FE}$ and thereby lead to a correction of the actual significance level towards the prescribed one. This procedure does not, in general, result in a conservative attitude of the test, as one may expect at first sight, because one has to keep in mind that the variance of the theoretical estimator $\tilde{\theta}_{FE}$ is estimated and not the variance of the feasible estimator $\hat{\theta}_{FE}$. Thus, even with an overestimation of the variance $\sigma^2(\hat{\theta}_i)$ in each study one may obtain too many unjustified significant results but in a smaller number than with the commonly used method (Böckenhoff & Hartung, 1998).

The above argumentation mainly holds if the variance estimator $1/\hat{v}$ underestimates the variances of the feasible estimator $\hat{\theta}_{FE}$. But if the variance estimator $1/\hat{v}$ overestimates the variance of $\hat{\theta}_{FE}$, the resulting test can be very conservative as Hartung and Knapp (2001) have observed, for example, in the log odds ratio case.

It is worthwhile to note that the just described deficiencies of the commonly used test statistic in the FE model are most striking if the sample sizes in the studies are small to moderate depending on the choice of the parameter of interest.

In the RE model, a similar result as in Theorem 4.5.4 can be stated for testing the hypothesis of no association using the test statistic T_2 from Equation 4.11. Suppose we use the untruncated unbiased estimator $\tilde{\tau}^2$ from Equation 4.8 of the between-study variance in T_2, then we can prove, following the lines of the proof of Theorem 4.5.4, that the variance estimator $1/\tilde{w}$, $\tilde{w}_i = [\tilde{\tau}^2 + \hat{\sigma}^2(\hat{\theta}_i)]^{-1}$, $i = 1, \ldots, k$, $\tilde{w} = \sum_{i=1}^k \tilde{w}_i$, in average underestimates the variance of the theoretical estimator $\tilde{\theta}_{RE}$. Thus, the test in the RE model may be rather anticonservative if the variance of the feasible estimator $\hat{\theta}_{RE}$ is near $1/w$. But on the other hand, the test can be very conservative if $1/\tilde{w}$ overestimates the variance of $\hat{\theta}_{RE}$. If the truncated estimator $\hat{\tau}^2_+$ from Equation 4.10 is used in the test statistic T_2 in practice, the expected value of the variance estimator is larger than the expected value of the variance estimator with the untruncated estimator $\hat{\tau}^2$ from Equation 4.9, but for growing τ^2 this difference diminishes.

4.6 AN ALTERNATIVE TEST STATISTIC IN THE FE AND RE MODEL

In Section 4.4, we have estimated the variances of the theoretical estimators $\tilde{\theta}_{FE}$ and $\tilde{\theta}_{RE}$ in the FE and RE model by estimating their components separately. Now we consider an estimator of the variance of $\tilde{\theta}_{RE}$ in the RE model 4.5 following the theory of variance components estimation (cf. Rao, 1972; Hartung, 1981). This estimator is a quadratic function of the study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, and is given in the RE model 4.5 as

$$\tilde{\sigma}^2(\tilde{\theta}_{RE}) = \frac{1}{k-1} \sum_{i=1}^{k} \frac{w_i}{w} \left(\hat{\theta}_i - \tilde{\theta}_{RE}\right)^2,$$

(cf. Hartung, 1999). Note that for $\tau^2 = 0$ the quadratic function $w \cdot (k-1) \cdot \tilde{\sigma}^2(\tilde{\theta}_{RE})$ coincides with the statistic Q from Equation 4.3, which is formally used in the FE model for checking the assumption of homogeneity.

Theorem 4.6.1. *The estimator $\tilde{\sigma}^2(\tilde{\theta}_{RE})$ is an unbiased estimator of the variance of $\tilde{\theta}_{RE}$ in the RE model 4.5.*

Proof. It holds

$$\begin{aligned} E\left(\hat{\theta}_i\right)^2 &= \operatorname{Var}\left(\hat{\theta}_i\right) + \left[E\left(\hat{\theta}_i\right)\right]^2 = w_i^{-1} + \theta^2, \\ E\left(\tilde{\theta}_{RE}\right)^2 &= w^{-1} + \theta^2, \\ E\left(\hat{\theta}_i\tilde{\theta}_{RE}\right) &= \frac{w_i}{w} E\left(\hat{\theta}_i\right)^2 + \sum_{j\neq i}^{k} \frac{w_j}{w} E\left(\hat{\theta}_i\right) E\left(\hat{\theta}_j\right) \\ &= w^{-1} + \theta^2. \end{aligned}$$

Then it follows

$$E\left(\hat{\theta}_i - \tilde{\theta}_{RE}\right)^2 = w_i^{-1} + \theta^2 - 2w^{-1} - 2\theta^2 + w^{-1} + \theta^2 = w_i^{-1} - w^{-1}$$

and

$$E\left(\tilde{\sigma}^2\left(\tilde{\theta}_{RE}\right)\right) = \frac{1}{k-1}\sum_{i=1}^{k} \frac{w_i}{w}(w_i^{-1} - w^{-1}) = \frac{1}{k-1}\left(kw^{-1} - w^{-1}\right) = \frac{1}{w}.$$

□

Moreover, it is shown in Hartung (1999) that the quadratic form $w \cdot (k-1) \cdot \tilde{\sigma}^2(\tilde{\theta}_{RE})$ is χ^2-distributed with $(k-1)$ degrees of freedom and stochastically independent of $\tilde{\theta}_{RE}$ in the RE model if the study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, are exactly normally distributed. Thus, we now consider the random variable

$$\left(\tilde{\theta}_{RE} - \theta\right) \Big/ \sqrt{\tilde{\sigma}^2(\tilde{\theta}_{RE})},$$

which under the above assumption is exactly t-distributed with $(k-1)$ degrees of freedom. Therefore, an alternative test of $H_0 : \theta = 0$ is given by

$$\frac{\left|\tilde{\theta}_{RE}\right|}{\sqrt{\tilde{\sigma}^2(\tilde{\theta}_{RE})}} > t_{k-1;1-\alpha/2}, \tag{4.12}$$

where $t_{\nu;\gamma}$ stands for the γ-quantile of the t-distribution with ν degrees of freedom.

The alternative test in 4.12, however, depends on the unknown between-study variance τ^2 and the unknown within-study variances $\sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$. Thus, for a practical application of this test we replace the unknown variance components by appropriate estimators. Then, the feasible test statistic reads

$$T_3 = \frac{|\hat{\theta}_{RE}|}{\sqrt{\hat{\sigma}^2(\tilde{\theta}_{RE})}} \tag{4.13}$$

and the hypothesis of no association is rejected if the observed value of T_3 exceeds the $(1-\alpha/2)$-quantile of the t-distribution with $(k-1)$ degrees of freedom.

4.7 COMBINED DECISION RULES

In the previous sections, we have always distinguished between the FE and the RE model. For practically conducting a meta-analysis one usually has to choose in advance between these two models. In the literature, there exist different opinions how to deal with this decision problem.

A widespread procedure is to make first a test of homogeneity using the test statistic Q_1 from Equation 4.6 and, if the hypothesis of homogeneity is rejected, one uses the RE model, otherwise the FE model (Normand, 1999). Note that the hypothesis of homogeneity in the FE model, that is, $H_0 : \theta_1 = \ldots = \theta_k$, is equivalent to the hypothesis that the between-study variance τ^2 in the RE model is equal to zero. But the test of homogeneity often has low power against the alternative $\tau^2 > 0$ so that one cannot satisfactorily avoid the false use of the FE model if a between-study variance is present. The effect of a dramatically increasing Type I error by using the test statistic T_1 from Equation 4.7, if heterogeneity is present, is, for example, shown in Ziegler and Victor (1999) and in our simulation study described in the next section.

Whitehead and Whitehead (1991) propose a similar decision making at first sight. They consider the method of moments estimator $\hat{\tau}^2$ from Equation 4.9 of the between-study variance and suggest to use the FE model if $\hat{\tau}^2$ yields a negative estimate, otherwise the RE model. But this procedure is identical to the principle to always use the RE model with the truncated variance estimator $\hat{\tau}^2_+$ from Equation 4.10.

Both procedures have in common that the decision for an analysis in a corresponding model depends on a judgement of the variation between the studies. If one is mainly interested in testing the hypothesis $H_0 : \theta = 0$, the "pre-test" determines the test procedure which has to be used, but a false decision of the "pre-test" may considerably affect the properties of the test procedure. The most crucial point in the just described choice between the two models is given if the true between-study variance τ^2 is relatively small. In this situation one may expect the most false decisions between the models. For growing τ^2 the decision for the RE model becomes more and more certain. Thus, we will consider decision rules for tests of the hypothesis $H_0 : \theta = 0$, which incorporate the test procedure in the FE model as well as in the RE model and depend only in part on a variance estimate of the between-study variance. Moreover, we make always use of the alternative test statistic T_3 from Equation 4.13 in the RE model.

The first combined decision rule is that the hypothesis $H_0 : \theta = 0$ is rejected if $T_1 > \chi^2_{1;1-\alpha}$ and $|T_3| > t_{k-1;1-\alpha/2}$, that means, we require that the commonly used test in the FE model as well as the alternative test in the RE model has to reject H_0. The idea behind this decision rule is that for a small between-study variance one may correct the significance level of the anticonservative test based on T_1 in the FE model towards the prescribed one, because the alternative test based on T_3 in the RE model may possess a smaller significance level. If the between-study variance grows, the role of the test based on T_1 becomes more and more ignorable and the combined decision rule is nearly identical to the decision rule simply based on the alternative test.

Furthermore, we consider two additional combined decision rules which include the estimation of the between-study variance. The first combined decision rule rejects $H_0 : \theta = 0$ if ($|T_3| > t_{k-1;1-\alpha/2}$ and $\hat{\tau}^2_+ > 0$) or ($T_1 > \chi^2_{1;1-\alpha}$ and $|T_3| > t_{k-1;1-\alpha/2}$, and $\hat{\tau}^2_+ = 0$), that is, we always require that the alternative test in the RE model rejects the hypothesis H_0 irrespective of the estimated value of τ^2, but if the truncated estimator $\hat{\tau}^2_+$ is equal to zero, that is, the usual method of moments estimator yields a negative estimate, the commonly used test in the FE model has also to reject H_0. This combined decision rule is motivated to correct a possible anticonservative attitude of the alternative test statistic in the RE model if the between-study variance is small and the test is anticonservative, while the commonly used test in the FE model is rather conservative in this situation. But if the alternative test statistic in the RE model performs better in case of small τ^2, this decision rule is nearly identical to the previous combined decision rule. Again, for growing τ^2 this decision rule becomes more and more similar to the decision rule based solely on the alternative test statistic T_3 in the RE model.

The second combined decision rule, which incorporates an estimation of τ^2, rejects $H_0 : \theta = 0$ if ($T_1 > \chi^2_{1;1-\alpha}$ and $\hat{\tau}^2_+ = 0$) or ($T_1 > \chi^2_{1;1-\alpha}$ and $|T_3| > t_{k-1;1-\alpha/2}$ and $\hat{\tau}^2_+ > 0$), that means, we always require that the commonly used test in the FE model rejects the hypothesis H_0 irrespective of the estimated value of τ^2, but if the truncated estimator $\hat{\tau}^2_+$ is greater than zero, the alternative test in the RE model has also to reject the hypothesis. This decision rule is similar to the proposal of Whitehead and Whitehead (1991), except that we use the alternative test statistic T_3 instead of the commonly used T_2 in the RE model. Since we also require to carry out the commonly used test in the FE model if the variance estimate of τ^2 is positive, this combined decision rule reduces the anticonservative attitude of this test for growing τ^2 as well as a possible anticonservative attitude of the test based on the alternative test statistic for small between-study variances.

4.8 SIMULATION STUDY

In a small simulation study, we compare the decision rules according to their actual Type I error rate. Table 4.1 summarizes all seven different decision rules we have already discussed and which are considered in the simulation study.

Table 4.1 Tests and Corresponding Decision Rules for Rejecting the Hypothesis $H_0 : \theta = 0$ at Level α With the Test Statistics T_1, T_2, and T_3

Test	Decision Rule: Reject $H_0 : \theta = 0$ if	
ψ_1	$T_1 > \chi^2_{1;1-\alpha}$	
ψ_2	$T_2 > \chi^2_{1;1-\alpha}$	
ψ_3	$\|T_3\| > t_{k-1;1-\alpha/2}$	

ψ_4	$T_1 > \chi^2_{1;1-\alpha}$,
	$T_2 > \chi^2_{1;1-\alpha}$,
		if $Q_1 \leq \chi^2_{k-1;1-\alpha}$
		if $Q_1 > \chi^2_{k-1;1-\alpha}$

ψ_5	$T_1 > \chi^2_{1;1-\alpha}$ and $\|T_3\| > t_{k-1;1-\alpha/2}$	

ψ_6	$\|T_3\| > t_{k-1;1-\alpha/2}$,
	$T_1 > \chi^2_{1;1-\alpha}$ and $\|T_3\| > t_{k-1;1-\alpha/2}$,
		if $\hat{\tau}^2_+ > 0$
		if $\hat{\tau}^2_+ = 0$

ψ_7	$T_1 > \chi^2_{1;1-\alpha}$,
	$T_1 > \chi^2_{1;1-\alpha}$ and $\|T_3\| > t_{k-1;1-\alpha/2}$,
		if $\hat{\tau}^2_+ = 0$
		if $\hat{\tau}^2_+ > 0$

Note. For a definition of T_1, T_2, and T_3 see Equations 4.7, 4.11, and 4.13, for Q_1 see Equation 4.6, and for $\hat{\tau}^2_+$ see Equation 4.10.

As an example, we choose the random one-way ANOVA model with heteroscedastic error variances, which is given by

$$y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \ldots, k;\ j = 1, \ldots, n_i, \tag{4.14}$$

with $a_i \sim \mathcal{N}(0, \tau^2)$ and $e_{ij} \sim \mathcal{N}(0, \sigma_i^2)$, and all random effects are stochastically independent. Instead of the individual data, usually summary statistics are given in publications. We consider the arithmetic mean $\hat{\mu}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ as the estimator of μ in each study. For this estimator, we have the following distributional property:

$$\hat{\mu}_i \sim \mathcal{N}\left(\mu, \tau^2 + \sigma^2\left(\hat{\mu}_i\right)\right), \quad \sigma^2(\hat{\mu}_i) = \sigma_i^2/n_i, \quad i = 1, \ldots, k. \tag{4.15}$$

Furthermore, an unbiased estimator of the error variance in model 4.14 is given by $\hat{\sigma}_i^2 = s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2/(n_i - 1)$, $i = 1, \ldots, k$, so that an unbiased estimator of the within-study variance $\sigma^2(\hat{\mu}_i)$ is $\hat{\sigma}^2(\hat{\mu}_i) = s_i^2/n_i$. This estimator is stochastically independent of $\hat{\mu}_i$ and it holds that $(n_i - 1)s_i^2/\sigma_i^2$ is χ^2-distributed with $(n_i - 1)$ degrees of freedom.

In the simulation study, we consider four different patterns of sample sizes and error variances, which are given in Table 4.2 for $k = 3$ studies. The first pattern has equal sample sizes and equal error variances, whereas in the second pattern the sample sizes are doubled in each study. In the last two patterns we consider different sample sizes in each study. In pattern 3, the error variances are increasing with growing sample sizes, but the within-study variance σ_i^2/n_i, $i = 1, 2, 3$, is always 0.1. In pattern 4, the error variances are decreasing with growing sample sizes so that the study with the largest sample size has the smallest within-study variance.

Table 4.2 Sample Sizes and Error Variances Used in the Simulation Study

Pattern for $k = 3$	Sample Sizes (n_1, n_2, n_3)	Error Variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$
1	$(5, 5, 5)$	(4,4,4)
2	$(10, 10, 10)$	(4,4,4)
3	$(10, 20, 40)$	(1,2,4)
4	$(10, 20, 40)$	(4,2,1)
Patterns for $k = 9$: Replicated Twice the Patterns for $k = 3$		

In Table 4.3, the results of our simulation study are put together. We present the results for $k = 3$ and $k = 9$ studies, where the patterns for $k = 9$ studies have been constructed by replicating the patterns for $k = 3$ studies twice. As different values of the between-study variance we choose $\tau^2 = 0, 0.1, 1, 10$, and as the estimator of τ^2 we always use $\hat{\tau}_+^2$ from Equation 4.10. Besides the estimated Type I error rates of the test given a prescribed significance level of $\alpha = .05$, the table also contains the estimated proportion of negative estimates of τ^2 using $\hat{\tau}^2$ from Equation 4.9 and the estimated power of the test of homogeneity based on Q_1 from Equation 4.6. Each estimated value in the table is based on 10,000 replications of the corresponding model.

From Table 4.3, we see that the commonly used test ψ_1 in the FE model is rather anticonservative in the normal mean case if the between-study variance is equal to zero, that is, the FE model is the theoretically correct one. Moreover, the Type I error rates increase if the number of studies grows, but if the sample sizes increase for fixed k the Type I error rates decrease. If between-study variation is present, the estimated Type I error rates become still larger and they increase for growing between-study variance. Consequently, one should be very careful in the normal mean case to apply the test ψ_1.

The commonly used test ψ_2 in the RE model, which coincides with the proposal of Whitehead and Whitehead how to deal with the choice between FE and RE model, yields its best results in comparison to the prescribed level if the between-study variance is equal to zero. From the estimated proportions of negative estimates of τ^2, we observe that for $k = 3$ in approximately 60 % of

Table 4.3 Estimated Type I Error Rates (in %) for the Seven Different Two-Sided Tests of $H_0 : \mu = 0$ in Model 4.15, Given $\alpha = .05$, the Estimated Proportion (in %) of Negative Estimates of τ^2, the Estimated Power (in %) of the Test of Homogeneity for $k = 3$ and $k = 9$ Studies

										$\hat{\tau}^2$	Power
k	Pattern	τ^2	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	neg.	(Q_1)
3	1	0	19.4	10.3	7.2	14.8	3.7	3.7	7.8	56.4	12.6
		0.1	21.3	11.0	7.2	15.9	4.2	4.2	8.4	53.2	14.6
		1	36.3	14.8	6.8	20.9	5.2	5.2	9.1	32.4	33.8
		10	69.3	18.6	5.1	20.5	5.0	5.0	5.8	6.3	82.7
	2	0	10.6	6.4	5.5	8.9	2.2	2.2	5.1	60.9	8.0
		0.1	14.8	8.2	5.9	11.7	2.8	2.8	6.2	51.7	13.4
		1	37.5	15.6	5.7	19.9	4.7	4.7	8.3	23.5	45.5
		10	73.7	18.7	5.2	19.3	5.2	5.2	5.7	3.7	89.4
	3	0	10.0	5.1	5.5	7.3	1.7	1.7	4.2	57.1	11.8
		0.1	26.4	13.7	7.7	20.1	5.2	5.2	10.7	44.4	20.6
		1	64.0	21.1	8.4	27.0	7.8	7.8	10.8	16.5	61.0
		10	87.7	21.0	6.1	21.5	6.1	6.1	6.2	2.2	93.7
	4	0	6.6	4.1	4.7	5.7	1.3	1.3	3.4	59.7	8.5
		0.1	35.2	16.6	9.1	25.2	6.8	6.8	12.3	38.7	24.9
		1	73.7	21.6	8.0	25.5	7.6	7.6	9.3	10.5	73.5
		10	91.1	22.2	5.8	22.5	5.7	5.7	5.8	1.4	96.0
9	1	0	26.9	9.8	9.4	15.3	7.6	7.7	8.1	32.9	29.8
		0.1	28.9	9.7	9.0	14.9	7.6	7.7	8.0	26.3	35.9
		1	44.5	9.8	6.8	12.8	6.7	6.7	6.7	5.5	74.7
		10	74.9	9.8	5.3	9.8	5.3	5.3	5.3	0.0	99.9
	2	0	12.1	6.5	6.6	9.3	4.6	4.6	5.2	44.7	14.1
		0.1	17.2	8.1	7.2	11.6	5.9	5.9	6.3	30.5	26.4
		1	39.7	8.6	5.4	10.0	5.3	5.3	5.4	2.2	86.2
		10	76.1	9.3	5.4	9.3	5.4	5.4	5.4	0.0	100
	3	0	11.2	5.4	5.3	7.5	3.7	3.7	4.1	40.8	19.5
		0.1	27.8	9.4	6.6	13.8	6.2	6.2	6.5	14.8	51.8
		1	66.7	10.3	6.0	10.7	6.0	6.0	6.0	0.3	97.7
		10	88.3	10.5	5.3	10.5	5.3	5.3	5.3	0.0	100
	4	0	7.5	4.9	5.4	6.6	3.3	3.3	3.8	49.1	11.4
		0.1	36.1	10.4	7.4	14.7	7.2	7.2	7.4	8.0	65.8
		1	73.9	10.0	5.5	10.0	5.5	5.5	5.5	0.1	99.6
		10	91.8	10.7	5.3	10.7	5.3	5.3	5.3	0.0	100

the simulated cases and for $k = 9$ in still approximately 40 % of the cases one actually performs the commonly used test in the FE model. If between-study variation is present, the estimated Type I error rates in the normal mean case increase to more than four times the prescribed level for $k = 3$ studies and to nearly twice the prescribed level for $k = 9$ studies. But this indicates that for an increasing number of studies the actual Type I error rate diminishes.

The alternative test ψ_3 in the RE model yields the best results concerning the actual significance level in comparison to the tests ψ_1 and ψ_2. If $\tau^2 = 0$ is not present, the tests ψ_2 and ψ_3 have rather similar estimated Type I error rates, but if a positive between-study variance exists, the alternative test has estimated Type I error rates often near the prescribed level and only in some cases goes beyond 7 %.

The test ψ_4, which has a decision rule depending on the test of homogeneity, always has an estimated Type I error rate which is greater or equal to the estimated Type I error rate of the test ψ_2. Thus, this combined decision rule does not yield an improvement concerning the actual significance level.

The test ψ_5, which requires the rejection of the hypothesis with the test ψ_1 and with the alternative test ψ_3, has always estimated Type I error rates which are less or equal to the estimated Type I error rates of the test ψ_3. The test ψ_5 is often an essential improvement in comparison to the test ψ_3 if the between-study variance τ^2 is equal to 0.1 or 1, but for $\tau^2 = 0$ the test ψ_5 may become rather conservative.

Due to the fact that the test ψ_1 is anticonservative the test ψ_6 yields nearly the same results as the test ψ_5, as we have pointed out in Section 4.7.

The test ψ_7 yields rather similar results like the test ψ_5 if $k = 9$ studies are considered. The estimated Type I error rates of ψ_7 are slightly greater or equal to the estimated Type I error rates of ψ_5. The relationship between these two tests also holds for $k = 3$ studies, but the difference between the estimated Type I error rates of the tests is much larger. Often, the estimated Type I error rates of ψ_7 are twice as large as the estimated ones of ψ_5.

REFERENCES

Böckenhoff, A., & Hartung, J. (1998). Some corrections of the significance level in meta-analysis. *Biometrical Journal*, *40*, 937–947.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Hartung, J. (1976). Some inequalities for a random function. *Teorija Verojatnostei i ee Primenenja*, *21*, 661–665 (See also: (1977) *Theory of Probability and its Application, SIAM, 21*).

Hartung, J. (1981). Nonnegative minimum biased invariant estimation in variance component models. *Annals of Statistics*, *9*, 278–292.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, *41*, 901–916.

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

Li, Y., Shi, L., & Roth, H. D. (1994). The bias of the commonly-used estimate of variance in meta-analysis. *Communications in Statistics – Theory and Methods*, *23*, 1063–1085.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*, 321–359.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, *67*, 112–115.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, *10*, 1665–1677.

Ziegler, S., & Victor, N. (1999). Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität [Some problems of the standard meta-analysis methods in the presence of heterogeneity]. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, *30*, 131–140.

5

Statistical Tests for the Detection of Bias in Meta-Analysis

Guido Schwarzer[1,3]
Gerd Antes[2,3]
Martin Schumacher[2]

[1] Freiburg Center for Data Analysis and Modelling
University of Freiburg

[2] Institute of Medical Biometry and Medical Informatics
University Hospital of Freiburg

[3] German Cochrane Centre
University Hospital of Freiburg

Summary

The result of a meta-analysis as part of a systematic review critically depends on the extent to which relevant information about the particular research question can be retrieved. Biases are especially to be expected due to the selective publication of significant results (publication bias).

For the investigation of biases in meta-analyses, both (informal) graphical as well as statistical methods are used. Within the framework of a simulation study, two tests for biases are compared; a rank-correlation test (Begg & Mazumdar, 1994) and a test based on a linear regression approach (Egger, Smith, Schneider, & Minder, 1997).

5.1 INTRODUCTION

The use of meta-analysis to combine the results of several independent trials is still increasing in the medical field. The validity of a meta-analysis may be affected by various sources of bias, for example, publication bias (Begg & Berlin, 1988; Egger, Smith, et al., 1997) and language bias (Egger, Zellweger-Zahner, et al., 1997). An analysis of bias should be a part of any systematic review. Both statistical tests and graphical methods have been proposed for this purpose. In this chapter, we describe these methods in some detail by the use of a simulated dataset with binary outcome data, which are common in medical applications.

Throughout the chapter, we utilize the following notation. Let t_i denote the estimated effect (e.g., the log relative risk or the log odds ratio) in trial i, $i = 1, \ldots, k$ with $E(t_i) = \mu$ and $\mathrm{Var}(t_i) = \sigma_i^2$. The estimated variance of t_i is denoted by v_i. The inverse variance method is used to derive an overall treatment effect

$$\bar{t} = \frac{\sum_{j=1}^{k}(t_j/v_j)}{\sum_{j=1}^{k}(1/v_j)},$$

where k is the number of trials involved in the meta-analysis (Cooper & Hedges, 1994).

We referred to a survey conducted at the German Cochrane Centre to generate sample sizes of individual trials. All issues from 1948 to 1998 of eight German medical journals were examined and information from all published primary randomized clinical trials was extracted (Galandi, personal communication). We fitted a log-normal distribution to this dataset restricted to trials with a total sample size of at least 30 patients. This log-normal distribution with mean 3.678 and variance 1.146 was used to generate sample sizes. A forest plot of 20 such generated trials with an underlying relative risk of 0.5 is displayed in Figure 5.1. The variance estimates v_i were calculated according to Fleiss (1993). Due to the small sample sizes, many trials in this specific meta-analysis have equal relative risk estimates, for example, five trials result in an estimated relative risk of 0.5. The estimated overall treatment effect is 0.542 with a 95% confidence interval [0.387; 0.757], which is in good agreement with the true treatment effect. This simulated dataset is used for illustrative purposes in the sequel.

5.2 GRAPHICAL METHODS FOR THE DETECTION OF BIAS IN META-ANALYSIS

A funnel plot is the most often used graphical method to check informally the presence of bias in meta-analysis. Beside this method, a radial plot can be used for this purpose, too.

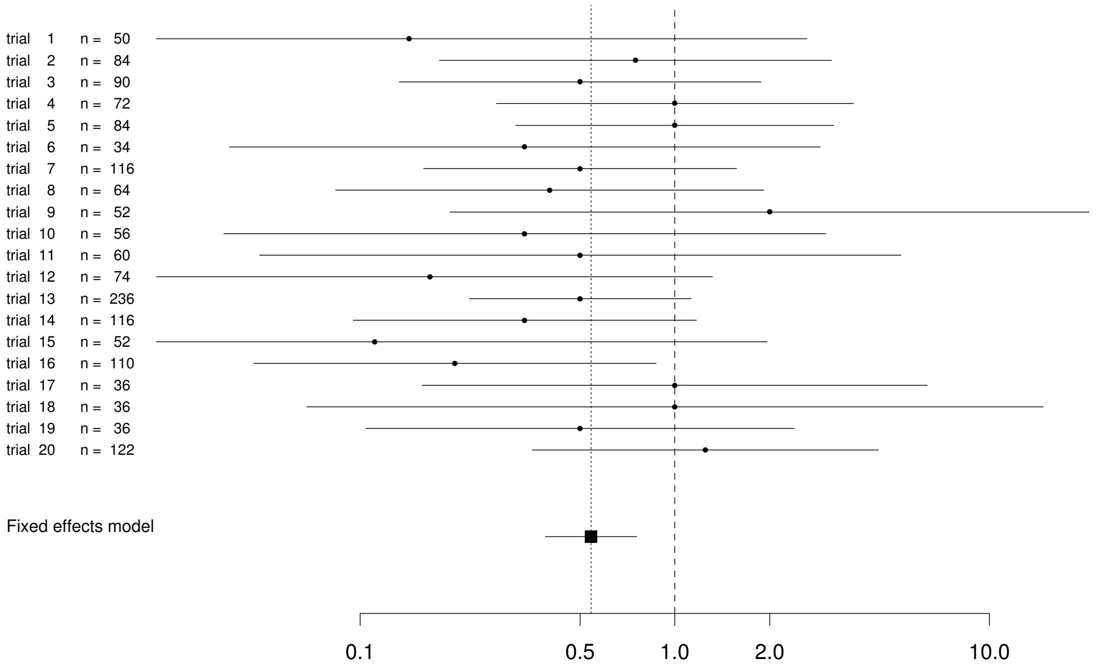


Figure 5.1 Forest plot of simulated meta-analysis; relative risk as measure of treatment effect.

5.3 FUNNEL PLOT

A scatterplot of the estimated treatment effect $x_i = t_i$ and a measure of the precision of t_i is called a funnel plot (Light & Pillemer, 1984). Typically, the sample size $y_i = n_i$ or the inverse of the estimated variance $y_i = 1/v_i$ is used as a measure of precision. For both measures, the display looks like a funnel if no publication bias and between-trial heterogeneity exists showing decreasing fluttering from bottom to top of the graph. Asymmetry in the funnel plot is taken as an indication of bias in the meta-analysis.

A variant of the funnel plot with standard error as measure of precision is displayed in Figure 5.2. The display should look like a triangle centered at the true treatment effect when the standard error is used as measure of precision. This kind of display has been chosen by the statistical methods group of the Cochrane Collaboration as the preferred variant.

We introduced a simple form of bias in the simulated meta-analysis as indicated by the plotting symbol in Figure 5.2. A funnel plot for the published trials (denoted by "s") can be derived from Figure 5.2 because the position of x_i and y_i which contain only trial specific information does not change. A meta-analysis considering only the published trials results in an estimated overall treatment effect of 0.385 with 95% confidence interval from [0.2546; 0.5821]. The asymmetry in the funnel plot is obvious if trials marked with "n" are not considered.

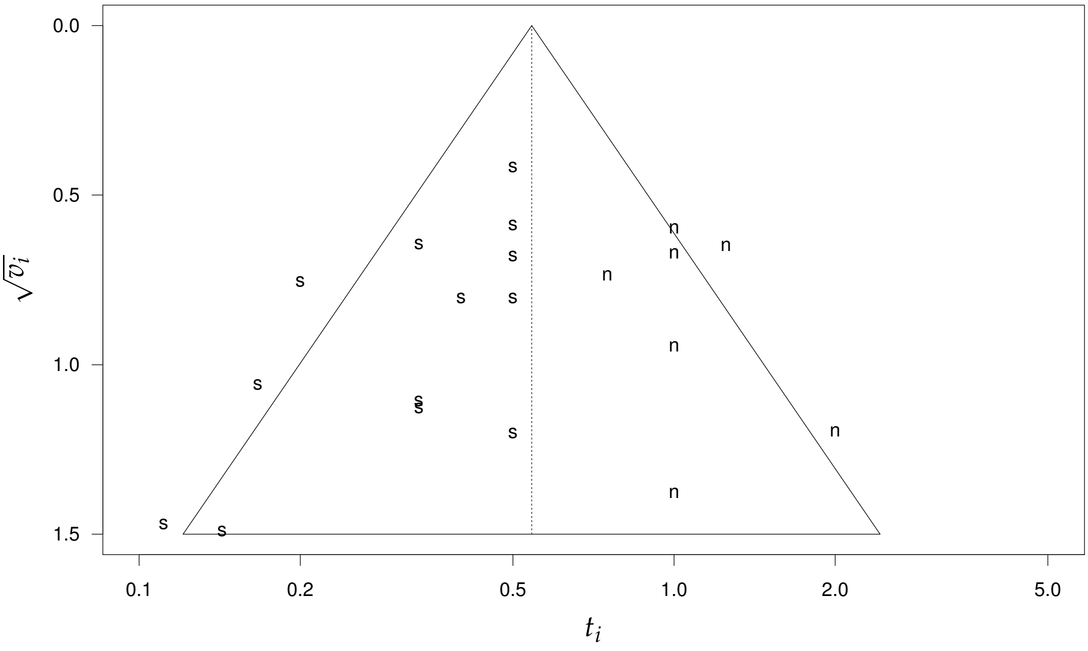


Figure 5.2 Funnel plot of simulated meta-analysis; relative risk as measure of treatment effect; s = trial published, n = trial not published.

5.4 RADIAL PLOT

Galbraith (1988b) introduced the radial plot in order to display several point estimates with different standard errors in a single graph. An additional paper focused on medical applications and the use of the log odds ratio as effect measure of interest (Galbraith, 1988a).

A scatterplot of $x_i = 1/\sigma_i$ and $y_i = t_i/\sigma_i$ is called a radial plot. A radial plot has the following properties (Galbraith, 1988b):

a) $\text{Var}(y_i) = 1$

b) t_i = slope of the line through (0,0) and (x_i, y_i)

c) Points are close to zero on the x-axis for large σ_i

d) Estimated overall effect $\bar{t}$ = slope in linear regression model: $y_i = \beta \cdot x_i + \epsilon_i$.

Due to properties b) and d), a circular scale is typically displayed on the right-hand side of a radial plot showing the estimated treatment effect. Sometimes $y_i^* = (t_i - \bar{t})/\sigma_i$ is plotted against x_i to get a better visual discrimination. In this case, the estimated overall effect coincides with the horizontal axis and departures from the overall effect are more obvious. In practice, the variances σ_i^2 are unknown and have to be estimated.

A radial plot of the simulated meta-analysis is depicted in Figure 5.3. Again, a plot for the published trials can be derived from this figure by omitting trials marked with n because the position of x_i and y_i does not change; a different

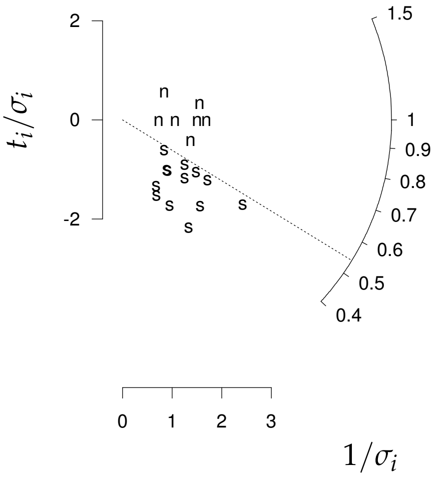


Figure 5.3 Radial plot of simulated meta-analysis; relative risk as measure of treatment effect; s = trial published, n = trial not published; overall treatment effect is indicated by the dashed line.

estimated overall treatment effect has to be considered for the published trials. A gap in the upper left part indicates the presence of bias if only published trials are considered. However, this is more obvious in the funnel plot.

5.5 STATISTICAL TESTS FOR THE DETECTION OF BIAS IN META-ANALYSIS

At least two test procedures for the detection of bias in meta-analysis enjoy some popularity. Begg and Mazumdar (1994) proposed a rank correlation test; Egger, Smith, et al. (1997) introduced a test based on a linear regression of the standard normal deviate on precision which is strongly connected to a radial plot. The estimated variance of the treatment effect in each single trial v_i is of central importance in both tests.

5.5.1 Begg and Mazumdar Test

Begg and Mazumdar (1994) proposed an adjusted rank correlation test for the detection of bias in a meta-analysis and evaluated the power of this test in a simulation study assuming a normal distribution for t_i. The test is based on the correlation between the standardized effect measure

$$t_i^* = \frac{(t_i - \bar{t})}{\sqrt{v_i^*}} \quad \text{with} \quad v_i^* = v_i - \frac{1}{\sum_{j=1}^{k} v_j^{-1}}$$

and the variance v_i. Kendall's tau is used as correlation measure. Let x denote the number of pairs of trials with standardized effects and variances ranked in the same order, that is, ($t_i^* > t_j^*$ and $v_i > v_j$) or ($t_i^* < t_j^*$ and $v_i < v_j$), where

$i \neq j$. The number of pairs ranked in the opposite order are denoted by y. The normalized test statistic for the case that no ties are present neither within t_i^* nor v_i is

$$z = \frac{(x-y)}{\sqrt{\frac{k(k-1)(2k+5)}{18}}},$$

where k is the number of trials involved in the analysis. A modified version for tied observations can be found in Armitage and Berry (1994). The test statistic z has an asymptotic standard normal distribution if the variances σ^2 are known.

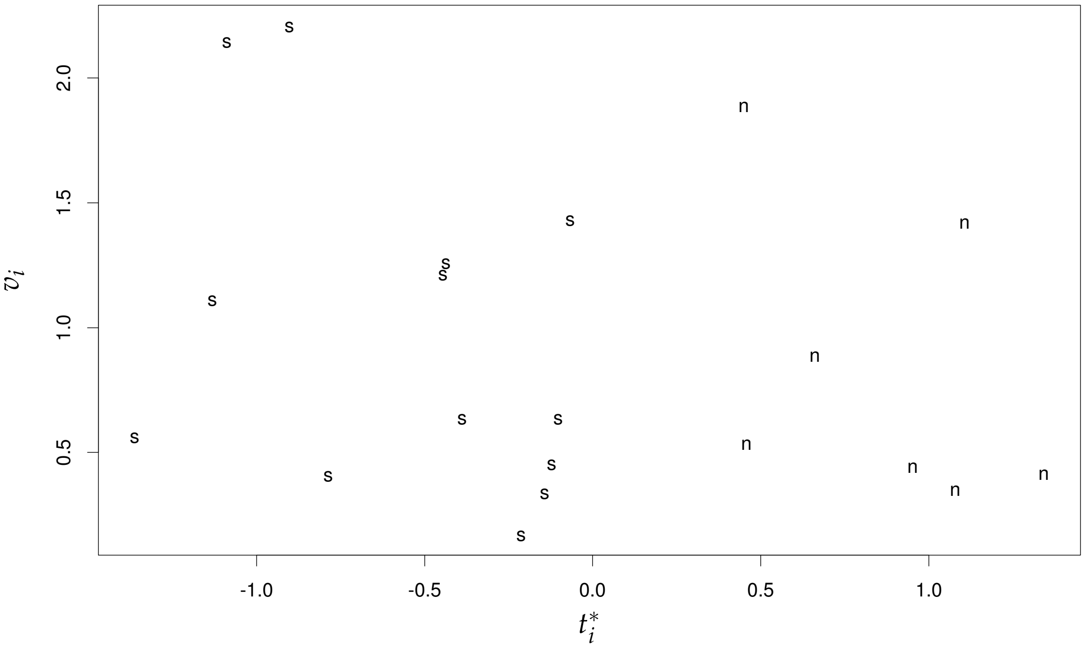


Figure 5.4 Graphical display of the rank correlation test for the detection of bias in the simulated meta-analysis; s = trial published, n = trial not published.

A scatterplot of t_i^* and v_i for the simulated dataset is depicted in Figure 5.4. No correlation between t_i^* and v_i is apparent if all trials are considered. This impression is supported by the result of the rank correlation test. The difference $x - y$ is -28 with a standard error of 30.8 resulting in a p-value of $p = .31$. The shape of the display for the published trials is different from Figure 5.4 because a different overall treatment effect $\bar{t}$ is utilized to calculate t_i^*. The difference is -30 with a standard error of 16.4 resulting in a p-value of $p = .067$ if only published trials are considered.

5.5.2 Egger Test

The test proposed by Egger, Smith, et al. (1997) for the detection of bias in meta-analysis is based on a linear regression of y_i on x_i: $y_i = \alpha + \beta \cdot x_i + \epsilon_i$. In contrast to the radial plot, the regression line is not constrained to run through the origin. A test for the detection of bias is constructed by testing the null-hypothesis of a zero intercept. The approach is justified by the intuitive argument that, in the presence of publication bias, small trials with non-significant or negative results are less likely to get published. Thus, points close to zero on the x-axis do not scatter randomly around the overall effect resulting in a non-zero intercept, that is, a departure from property d) of a radial plot. The test procedure is implicitly based on the assumption that linearity still holds in the presence of bias.

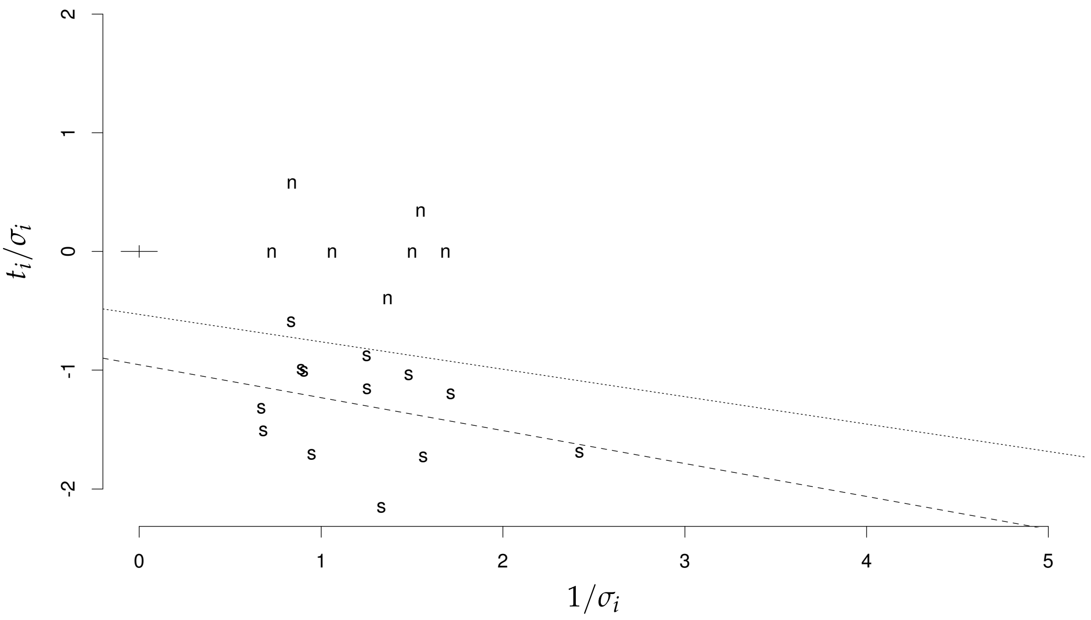


Figure 5.5 Result of the Egger test for the detection of bias in the simulated dataset; x_i and y_i according to a radial plot; regression lines displayed both for all trials and subset of published trials; s = trial published, n = trial not published.

The result of the Egger test for the simulated meta-analysis is displayed in Figure 5.5. The estimated intercept, if all trials are considered, is $\widehat{\alpha} = -0.53$ with a standard error (SE) of $SE(\widehat{\alpha}) = 0.532$ compared to a t-distribution with $18\,df$, resulting in a p-value of .33. A clear indication of bias is given if only published trials are considered: $\widehat{\alpha} = -0.95$ with $SE(\widehat{\alpha}) = 0.33$ resulting in a p-value of .015 (compared to a t-distribution with 11 degrees of freedom).

5.6 CONCLUDING REMARKS

In this chapter, we described two statistical tests on bias in meta-analysis which have been developed recently. Both tests are implicitly based on the assumption that the variances σ_i^2 are known. The statistical properties of these tests in practical relevant situations are still unknown. Further research is needed, especially with regard to the usefulness in meta-analysis with binary outcome data and in the case of heterogeneity.

We utilized a very simple method to generate bias in meta-analysis by arbitrarily omitting trial results. In order to compare the tests on bias in meta-analysis in simulations, more sophisticated mechanisms to generate bias are needed. This simulation model could be based on an approach described in Copas (1999) linking the probability of trial publication to both sample size and magnitude of observed treatment effect.

REFERENCES

Armitage, P., & Berry, G. (1994). *Statistical methods in medical research* (3rd ed.). Oxford: Blackwell.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A*, *151*, 419–445.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Copas, J. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A*, *162*, 95–109.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Egger, M., Zellweger-Zahner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *The Lancet*, *350*, 326–329.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, *2*, 121–145.

Galbraith, R. F. (1988a). A note on estimated odds ratios from several clinical trials. *Statistics in Medicine*, *7*, 889–894.

Galbraith, R. F. (1988b). Graphical display of estimates having differing standard errors. *Technometrics*, *30*, 271–281.

Light, R. J., & Pillemer, D. B. (1984). *Summing up. The science of reviewing research.* Cambridge: Harvard University Press.

6

Issues of Traditional Reviews and Meta-Analyses of Observational Studies in Medical Research

Wilhelm Sauerbrei

Institute of Medical Biometry and Medical Informatics
University Hospital Freiburg

Maria Blettner

Department of Epidemiology and Medical Statistics
School of Public Health
University of Bielefeld

Summary

Summarizing data from several studies is an important part in medical research. Several problems of traditional review articles are known for a long time, with the consequence to demand for more systematic reviews. We will outline the rationale for meta-analyses and describe four methods to summarize data, with the emphasis on observational studies where the association of risk or prognostic factors and certain diseases are investigated. We will compare and assess several criteria for different types of overviews such as narrative review, meta-analysis from literature, meta-analysis with individual patients data, and the prospectively planned meta-analysis. We will critically discuss some examples from the literature and will show severe problems of meta-analyses based on literature data only. We argue that a reasonable and valid meta-analysis of observational studies requires in general some re-modeling of the data.

Therefore, the use of individual data is an important requirement to reach reliable conclusions on the association between the factor and the outcome of interest.

6.1 INTRODUCTION

The serious problems and questionable recommendations from traditional review articles have been shown and the necessity of more systematic reviews in a timely fashion by using statistical techniques are well known (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992). Therefore, much attention has been given in recent years to meta-analysis in medical research, however, numerous methodological issues particularly with respect to biases and the use of meta-analysis are still raising controversial discussions (Chalmers, 1991; Chalmers & Lau, 1993; Thompson & Pocock, 1991; Stewart & Parmar, 1993). Authors have heavily criticized the method as such ("If a medical treatment has an effect so recondite and obscure as to require meta-analysis to establish it, I would not be happy to have it used on me", Eysenck, 1994, p. 792) or identified poorly performed meta-analyses ("In my own review of selected meta-analyses, problems were to frequent and so serious, including bias on the part of the meta-analyst, that it was difficult to trust the overall 'best estimates' that the method often produces.", Bailar, 1997, p. 560), both resulting in some discredit of this method. Additionally, in many circumstances such as in medical decision making where modern techniques of health technology assessment (HTA) play a central role, often estimates of parameters are needed and produced by a "quick" meta-analysis. Deficiencies in the meta-analysis may transfer to unsound decisions.

The critique of meta-analysis should distinguish between three central aspects:

1. A major distinction should be made whether the results of randomized trials (RCT) or observational studies are summarized. Studies comparing a treatment given in one clinic with another treatment given in another clinic are here seen as observational studies. As parts of the experiment are under control of the investigators, these studies are often considered as quasi-experiments.

2. A second important feature is the measurement scale of the factors of interest. Different problems occur depending on whether binary, ordinal, categorical, or continuous factors are investigated. Additional problems occur with a summary assessment using meta-analytic techniques when different scales of measurement were used in the original studies.

3. The third and most often considered characteristic is whether individual patient data (often called MAP= Meta-Analysis of Patient) or published data (called MAL = Meta-Analysis of Literature) is used for the meta-analysis.

There is general acceptance that meta-analyses of RCT based on individual data give the most reliable results concerning the combination of studies. Obviously, conducting a MAP is a large study in its own which requires large effort and funds from the group starting such a project and the willingness of the investigators from the single studies to cooperate. For a detailed discussion, see Stewart and Clarke (1995). These difficulties are usually given as the main justification for a meta-analysis based on published data only. Concerning RCTs, well-conducted meta-analyses based on MAL can result in useful summaries of an effect of interest and they are an accepted instrument. For observational studies the situation is more complex as the studies are less homogeneous, for example, adjustment for confounder factors in a multivariate model is essential for each single study and those are often different between studies.

In this chapter, we will discuss merits, limitations, and difficulties of different types of systematic reviews for observational studies. The following is partly taken from a paper by Blettner, Sauerbrei, Schlehofer, Scheuchenpflug, and Friedenreich (1999), where the topic in epidemiology is considered and terminology from this field is often used. Most arguments given in this paper apply to other observational studies in clinical research.

Four different methods for summarizing the evidence are distinguished:

Review: Qualitative summary, the narrative review article.

MAL: Meta-Analysis of literature, that is, quantitative summary of published data.

MAP: Meta-Analysis of patient data, that is, re-analysis of individual data of the original studies followed by a quantitative summary.

Prospective MA: Prospectively planned, pooled analysis of several studies, where pooling is already a part of the protocol. Data collection procedures, definition of variables, questions, and hypotheses are as far as possible standardized for the individual studies.

In the literature, different terms are used for these four types and certainly some combinations are possible. It should be noted that the prospective meta-analysis differs in several respects from a classical multicentre clinical trial since often the single studies are analyzed and published separately. In many situations, the design of the studies is slightly different because of local or regional circumstances.

For the following, we summarize and compare the four different types for investigations where the influence of one or several factors on an outcome variable is investigated in non-randomized studies. First, we give the major reasons and some general rules for conducting a meta-analysis. Then we describe the similarities and differences between the various types. We compare the advantages and limitations of the four types. We will discuss some examples of summaries of observational studies based on published data for risk factors, prognostic factors, and therapeutic factors mainly with the emphasis to demonstrate severe problems of meta-analyses from literature.

6.2 RATIONALE FOR META-ANALYSES

The main reason for conducting a review or a meta-analysis is to summarize the results of previously conducted studies which usually have inconsistent results. Such a situation may arise when the sample sizes of individual studies are too small to find stable results or if the results from single studies vary considerably. Meta-analyses are mainly used to assess the influence of weak risk factors, which may nevertheless have a large public health impact (such as passive smoking, use of contraceptives, or exposure to electromagnetic fields) or of treatment strategies whose small benefits can be worthwhile for a severe disease with a large incidence. For other issues in medical research, for example, prognostic factor studies or diagnostic studies, the use of meta analyses is increasing. Review articles investigate whether the available evidence is consistent and/or to which degree inconsistent results can be explained by random variation or by systematic differences between the design, the setting or the analysis of the study. In contrast to qualitative reviews, MAL or MAP are mainly performed to obtain a combined estimator of the quantitative effect of the risk factor such as the relative risk or risk difference. Some meta-analyses are also used to investigate more complex dose-response functions. The majority of meta-analyses conducted so far examined dichotomous (or categorical) factors. Briefly, the main reasons for conducting a meta-analysis or a review of observational studies are:

1. to assess qualitatively whether a factor has to be considered as a risk factor,
2. to provide more precise effect estimates and increased statistical power and to analyze dose-response relations,
3. to investigate the heterogeneity between different studies,
4. to generalize results of single studies,
5. to investigate rare exposure and interactions, and
6. to investigate risks associated with rare diseases.

6.3 CHARACTERIZATION AND LIMITATIONS OF THE FOUR TYPES

6.3.1 Review

Traditional narrative reviews provide a qualitative but not a quantitative assessment of published results. They are influenced by publication bias (Dickersin, 1990, 1997) and the file drawer problem (Rosenthal, 1979). If there is not an *a priori* strict protocol for the review, narrative reviews are only a subjective judgment of the included studies. However, if they are carefully done, they can give quite an extensive overview of the current state of the research within a short time frame and at low cost.

6.3.2 Meta-Analysis From Literature (MAL)

These studies are comparable to a narrative review with respect to time and cost, with the main difference being that the primary goal is to give a quantitative estimate of the effect of interest. They can be performed from published data without cooperation and without the agreement of other study groups. However, attempts may be made to obtain additional information from study coordinators, if necessary. So far, not many meta-analysts have tried this approach. There are some major limitations of this approach that have been pointed out by several authors (see e.g., Shapiro, 1994). One limitation is that publication bias is particularly important since some explorative analyses may be done and published selectively. Most likely, unexpected significant results may be selected for publication, yielding an overestimation of the effect. An additional problem is that studies may differ considerably in designs, data collection methods, and the precise definition of the factors of interest and the confounder variables. A special dilemma arises if different studies adjust for different confounding factors. No systematic investigation has been performed to determine whether the simple (crude) estimates or "best estimates" should be used for combining results of individual studies. Many aspects of the heterogeneity cannot be dealt with appropriately in such summaries. A combined estimate should not be calculated if the heterogeneity between studies is too high. However, in many publications, the problem of heterogeneity is not adequately handled. An estimate is often published although strong heterogeneity between study results was observed.

6.3.3 Meta-Analyses With (Individual) Patients Data (MAP)

Some of the problems that arise with MAL are avoidable if individual data from all investigations performed on the subject matter are available. Publication bias may be less prominent as it is possible that investigators are willing to contribute their data even if for a specific theme no analysis has been performed and no papers have yet been published. The cooperation between different researchers may help to identify studies, for example, if they are only known to local investigators. With individual data, statistical re-analysis can be performed. This analysis can include a new unified definition of the available variables and new regression models. With a large number of patients the effect of rare exposures can be examined. New hypotheses for specific subgroups may also be investigated.

It is often argued that major barriers for MAP are the high cost and long duration, and that it requires close cooperation between researchers (Stewart & Parmar, 1993; Stewart & Clarke, 1995). Although an improvement of data quality is not possible, some errors in the data or in the statistical analysis can be corrected for. Furthermore, adjustment for confounding variables that have been delineated since the original studies were performed can be done if those covariables were originally collected. Differences between the study results can be actively discussed between study coordinators and reasons for these differences can be elucidated. In general, it is possible to estimate risk coefficients and their variance from the combined data.

6.3.4 Prospectively Planned Meta-Analysis

This type of analysis has not been called a meta-analysis, despite the fact that it has several aspects in common with MAL and MAP. Several large international case-control studies and occupational cohort studies have used this approach (e.g., Boffetta et al., 1997). The major difference is that joint planning of the data collection and analysis makes it possible to avoid large differences between the studies since many details can be planned in advance and standardized. The experience of many coordinators is used in the preparation of the new multicentre study to ensure comparability in design, data collection, data analysis, and reporting across all centres. In contrast to multicentre randomized clinical trials, more heterogeneity in the individual study centres may exist arising from differences in populations (e.g., race is not a confounder factor in Germany but in the United States) or in design (e.g., no methods of random population sampling exist in the U.S., no overall cancer registration in Germany). The costs for a new multicentre study are in general high. The planning phase can be substantial, even difficult, and the time incurred can be long. Alternatively, individual studies with a joint core protocol for questions of common interest may be performed. This allows individual researchers to set priorities and also permits some variation across studies. One disadvantage of a large multicentre study is that errors in the design can be multiplied. Moreover, once a meta-analysis has been performed it will be more difficult to justify a new individual study with the same topic.

6.4 METHODS FOR AN OVERVIEW

All types of overviews – whether quantitative or qualitative – have some steps in common that should be followed in planning and conducting it. Each individual type has some aspects of the conduct that are different and that will be described later.

6.4.1 Steps in Performing a Meta-Analysis

Each type of overview needs a clear study protocol that describes the research question and the design, including how studies are identified and selected, the statistical methods to use, and how the results will be reported. This protocol should also include the exact definition of the disease of interest, the factors of interest, and the potential confounding variables that have to be considered. A main component of the protocol is the exact definition of the inclusion criteria for single studies. As described by Friedenreich (1993), the following steps are needed for a meta-analysis:

1. Define a clear and focused topic for the review.
2. Locate all studies (published and unpublished) that are relevant to the topic.
3. Select all studies that are relevant according to the explicit inclusion criteria.
4. Abstract necessary information from the published papers or obtain the primary data from the original investigators. Meta-analysis of published data may also include contacting the original project leaders to obtain data or information that have not been published in sufficient detail. For a MAL, agreement to use the original data is needed.
5. Tabulation of relevant elements of each study, including sample size, assessment procedures, available variables, study design, publication year, performing year, geographical setting, and so forth.
6. Define protocol for the analysis of all studies and estimate the study-specific effects (relative risks adjusted for relevant confounder variables).
7. Investigate the homogeneity of study-specific effects and determine whether these effects can be combined to perform a pooled analysis.
8. Presentation of published results, for example, graphically.
9. Investigate and reduce (if possible) the heterogeneity between studies.
10. Decide about remaining heterogeneity components: Coping with different designs, study types, confounder, and so forth.
11. Estimate a pooled effect with adequate statistical methods if the studies are efficiently homogeneous.
12. Conduct a sensitivity analysis.

Obviously, some variations are needed for the different forms, for example, for traditional reviews, in general, only the steps 1 to 4 together with a qualitative assessment are done. For a meta-analysis from published data, data abstraction will be done from publications, however, if required data are not given in the publications, contact with the project manager of the studies should be made.

6.4.2 Statistical Analysis

The statistical analysis of aggregated data from published studies was first developed in the fields of psychology and education (Glass, 1977; Smith & Glass, 1977). These methods have been adopted since the mid-1980s in medicine primarily for randomized clinical trials and are also used for observational studies. We will give a brief outline of some issues of the analysis. For more details we refer to several textbooks or to a recent tutorial paper (Normand, 1999).

6.4.2.1 Single Study Results A first step of the statistical analysis is the description of the characteristics and the results of each study. Tabulations and simple graphical methods should be employed to visualize the results of the single studies. Plotting the odds ratios and their confidence intervals (so-called forest plot) is a simple way to spot obvious differences between the study results. The Galbraith plot (Galbraith, 1994) is a more sophisticated way to investigate the heterogeneity and the contribution of each study to the overall estimate.

6.4.2.2 Heterogeneity The investigation of the heterogeneity between the different studies is a main task in each review or meta-analysis (Thompson, 1994). For the quantitative assessment of heterogeneity, several statistical tests are available (Petitti, 1994; Paul & Donner, 1989). A major limitation of formal heterogeneity tests is, however, their low statistical power to detect any heterogeneity present. Therefore, it is recommended to investigate the heterogeneity informally, for example, by comparing results from studies with different designs, within different geographical regions. In addition, graphical methods should be used to visualize heterogeneity, such as plots with single studies grouped or ordered according to special covariables as type of study, publication time, etc., or funnel plots to indicate publication bias, and radial plots (Galbraith, 1994).

In meta-analysis of literature (MAL), some sensitivity analysis can be performed to investigate the degree of heterogeneity. However, if individual data are available, the sources of heterogeneity can be investigated in some detail. Heterogeneity can be reduced, for example, by using the same statistical model for all single studies. In a prospective meta-analysis, the strategy for the statistical analysis and the definitions of variables can be determined *a priori* for all individual studies. Hence, an identical multiple regression analysis can be used in each centre. This avoids heterogeneity that could be introduced by different models.

6.4.2.3 Summarizing Effect Estimates Whether calculating a common estimate is appropriate should be decided after investigating the homogeneity of the study results. If the results vary substantially, no estimator should be presented or only estimators for selected subgroups should be calculated (e.g., combining results from case-control-studies only). Methods for pooling depend on the data available. In general, a two-step procedure has to be applied.

First, the risk estimates and variances from each study have to be abstracted (MAL) or calculated (MAP). Then, a combined estimate is obtained as a (variance based) weighted average of the individual estimates. The methods for pooling based on the 2 × 2 table include the approaches by Mantel-Haenszel and Peto (see Petitti, 1994, for details). If data are not available in a 2 × 2 table but as estimates from a more complex model (such as an adjusted relative risk estimate), the Woolf and DerSimonian-Laird approach can be adopted using the estimates and their (published or calculated) variance resulting from the regression model (DerSimonian & Laird, 1986). For these methods, variance estimates of the pooled estimator are available and allow the calculation of confidence intervals.

Usually two different statistical concepts are used for the combined estimator. In *the fixed effects model,* it is assumed that the underlying true exposure effect in each study is the same. The overall variation and, therefore, the confidence intervals will reflect only the random variation within each study but not any potential heterogeneity between the studies. If individual data is available, the pooled estimator and its variance can be obtained using regression models by incorporating an additional dummy variable for each centre. The *random effects model* incorporates variation between the studies. It is assumed that each study has its own (true) exposure effect and that there is a random distribution of these true exposure effects around a *central* effect. The observed effects from the different studies are used to estimate this distribution. In other words, the random effects model allows non-homogeneity between the effects of different studies.

The most common approach to combine the single estimates is the methods of moments given by DerSimonian and Laird (1986). The important difference is that for this model, study specific weights are calculated as a sum of the variance within the studies and a term for the variance between the studies, τ^2. The between-study variance τ^2 can also be interpreted as a measure for the heterogeneity between studies. Because of anticonservatism in case of the validity of a random effects model, Ziegler and Victor (1999) proposed a modification of the test based on DerSimonian and Laird (1986). The new proposal holds the nominal level asymptotically.

Comparison between fixed effects and random effects model

- Random effects methods yield (in general) larger variance and confidence intervals than fixed effects models because a between-study component τ^2 is added to the variance.
- If the heterogeneity between the studies is large, τ^2 will dominate the weights and all studies will be weighted more equally (in random effects model weight decreases for larger studies compared to the fixed effects model)
- A major critique of the random effects model is that it is not sufficient to "explain" the heterogeneity between studies, since the random effect merely quantifies unexplained variation by estimating it (e.g., Mengersen,

> Tweedie, & Biggerstaff, 1995). Heterogeneity between studies should yield careful investigation of the sources of the differences. If a sufficient number of different studies are available, further analyses, such as "meta-regression", may be used to examine the sources of heterogeneity (Greenland, 1987, 1994).

If individual data is available, the fixed effects estimate can be calculated from a regression model with dummy variables. So far, there is no comparable approach available for the random effects model. Here, the two-step procedure is used even with individual data available (e.g., Lubin et al., 1995).

Several other methods have been proposed to estimate the overall effect based on maximum-likelihood methods or on Bayesian methods (DuMouchel, 1990; Smith, Spiegelhalter, & Thomas, 1995). Recent investigations have demonstrated that, for practical purposes, the differences between these methods are not very large. So far, only rather sophisticated software is available for these approaches (Spiegelhalter, Thomas, Best, & Gilks, 1996).

6.4.2.4 Sensitivity Analysis An important method for investigating heterogeneity is sensitivity analysis, for example, to calculate pooled estimators only for subgroups of studies (according to study type, quality of the study, period of publication, etc.) to investigate variations of the odds ratio. An extension of this is meta-regression as proposed by Greenland (1987), however, this method cannot be used in most meta-analyses since too few studies are available.

6.5 COMPARISON AND ASSESSMENT OF THE FOUR TYPES OF REVIEWS

The different review methods are outlined in Table 6.1 and will be discussed here in detail.

6.5.1 Design, Conduct and Literature Search

For each type of review, the hypothesis, question, and conduct should be summarized and defined in a strict protocol in which clear inclusion and exclusion criteria for the studies and the details of the literature search are described. This component of the review process is important for each study type, but is especially needed if quantitative results are required. It should also be decided whether and which data will be required from the investigators of the individual studies.

An important problem of meta-analysis is publication bias. This bias has received a lot of attention particularly in the area of clinical trials. Publication bias occurs when studies that have non-significant or negative results are published less frequently than positive studies. For randomized clinical trials, it has been shown that even with a computer-aided literature search only some of the relevant studies will be identified (Dickersin, Scherer, & Lefebre, 1994). For observational studies additional problems exist: Very often a large number

Table 6.1 Comparison of Methods for Different Literature Review Methods

Requirement for the Review Method	Review	MAL	MAP	PMA
Planning and literature search				
Protocol	+?	+	++	++
Inclusion / Exclusion criteria	+	+	++	++
Systematic literature search (incl. Abstracts, Proceedings)	+?	+	++	$*$
Obtaining additional information from single studies that are not published	−	+?	+	$*$
Evaluation of sources of errors and bias				
Investigation of sources of bias	+?	+?	++	++
Evaluation of validity of individual studies	−	+?	++	$*$
Control of data collection	−	−	+?	++
Adjustment of inclusion criteria for individuals	−	−	+	++
Assessment and control of statistical analysis	−	−	++	++
Estimation of publication bias	−	?	+	$*$
Comparability of single studies				
Standardized study design	+?	+	+	++
Standardized assessment of risk factors	−	−	−	+
Standardized definition of exposure and confounder variables (categories)	−	−	+?	++
Standardized adjustment for confounder variables	−	−	+?	++
Statistical analysis				
Quantitative estimate for the effect	−	+?	++	++
Improvement of the precision of effects measured	−	?	+	++
Estimator for dose-response relationship	−	−	+?	++
Estimator for risk in subgroups	−	?	+	+
Increase of statistical power	−	+?	++	++
Evaluation of interactions and confounder effects	−	−	+	++
Evaluation of sources of heterogeneity	?	+?	++	++
Sensitivity analyses	−	+	++	++
Reproducibility of methods	−	−	+?	+?
General aspects				
Description of state of research	+	+	+	$*$
New research questions	++	++	+	+
Improvement of the quality of further studies	+	+	+	+
Time and costs for the study	very low	low	high	very high

Note. MAL = Meta-analysis of literature, MAP = Meta-analysis with patient data, PMA = Prospective meta-analysis, ++ = possible in principle and (almost) always done, + = possible in principle and often done, +? = often not possible or not done, ? = only possible or useful in exceptional cases, − = never possible, $*$ = less relevant.

of variables will be collected in questionnaires as potential confounders. If one or several of these potential confounders yield significant or important results, they may be published in additional papers, papers that have often not been planned in advance. If these confounders, however, yield expected or negative results, no publication will be made. Some regional studies may not be published in international journals and are not found in a literature search for a meta-analysis.

Inclusion criteria, data collection methods, and statistical analyses cannot be changed if published data are used for a meta-analysis. In many situations it is even difficult to determine exactly what has been done from the published literature. The methods section in many papers is often short and critical evaluation is not always possible. Errors in the original work cannot be corrected or checked and may yield to bias in the results of the meta-analysis. For MAP, the inclusion criteria for the single studies can be modified (for different age groups, tumour sites, latency times, etc.). They can also be redefined and checked. It is also possible to evaluate or adopt the statistical analysis if patient data is available. Possible sources of a systematic bias can be eliminated if a detailed statistical analysis of the single studies can be done. The evaluation of possible bias attributable to lack of control for confounding is also only possible with individual data.

6.5.2 Validation of Comparability of the Single Studies

Since many study designs are possible, it is necessary to evaluate the comparability of the single studies before conducting a review. This evaluation can be conducted partly from published data if enough detailed information is available in the papers. If individual data are available, an analysis of the single studies in one common model is possible. A major reason for different results across studies is that different statistical methods/models have been used. Hence, heterogeneity can be significantly reduced in a pooled analysis by using the same model for all studies. A pooled analysis is only possible if similar data are available from all studies and are provided to the investigator of the pooled analysis. The investigation of study-specific heterogeneity can be done, to some extent in MAL, mainly with a sensitivity analysis.

6.5.3 Quantitative Risk Estimation

Reviews are not designed to give a quantitative estimate of the effect of risk factors or to describe a dose-response relation. They only allow a descriptive comparison of the results of available research. All other types of meta-analysis allow – if the data are sufficiently homogeneous – the calculation of a pooled risk estimate. A quantitative estimate is very often considered an important goal of meta-analysis. However, calculating an overall risk estimate is not always possible because different statistical models were used in the original studies that prohibit a sensible combination of study results to be made. It should also be noted that an improvement in the precision of the risk estimate could often not be achieved by pooling since only the variation caused by random error (increasing the sample size) will be decreased by pooling. Increasing the sample size cannot eliminate any bias (systematic error). Indeed in some situations, the pooled estimate is less precise than estimates of the included single studies, as was shown by Gilbert (1989) in the radiation leukaemia studies.

A less precise estimate is likely if only data for a crude categorization (e.g., 2×2 tables) can be abstracted from the publication. Bias may also be increased if different methods to control confounding have been used in the individual studies. A more precise estimate is in most circumstances only possible in a re-analysis with individual patient data. Especially, to estimate a dose-response relation, individual patient data are required at least if different categories are used. Likewise, an investigation of interaction and confounding requires individual data. Prospective multicentre meta-analyses have the advantage that the data collection procedures, the measurement methods, the exposure assessment, and the definition of all variables can be agreed upon prior to data collection. Consequently, the data can be more easily combined at a later stage. It should also be noted that subgroup analysis, which is often a goal of a planned meta-analysis, could only be performed if the data are published with sufficient detail.

To investigate whether the results are consistent across studies, published data can be used for a review as well as for MAL. However, only limited search for sources of this heterogeneity is possible. For example, whether different definitions of exposure and confounder variables or different use of confounder in a multivariate statistical model influence the results can not be determined. A valid judgement of the consistency of results in complex questions requires a new and detailed statistical analysis based on original data.

Many authors have pointed out that investigating heterogeneity is the most important aspect of meta-analysis (e.g., Thompson, 1994). Statistical methods to investigate heterogeneity can be based on aggregated data. However, statistical tests have low power and may not be able to detect heterogeneity between studies. MAP allows different strategies to be used to eliminate differences and at least to give results in a unified way. Frequently, it is difficult to compare results from different observational studies since different data presentation methods are used across publication. Even in a single study different strategies for modeling can yield rather different results (Blettner & Sauerbrei, 1993). Therefore, for a meaningful meta-analysis it is necessary to eliminate this source of heterogeneity. Such a comparison is only possible if the same (or quite similar) variables are available from all studies.

6.6 SOME EXAMPLES

Ursin, Longenecker, Haile, and Greenland (1995) report results from a meta-analysis investigating the influence of the Body-Mass-Index (BMI) on development of premenopausal breast cancer. They include 23 studies of which 19 are case-control studies and 4 are cohort-studies. Some of these studies were designed to investigate BMI as risk factor, others measured BMI as confounders in studies investigating other risk factors. It can only be speculated that the number of unpublished studies in which BMI was mainly considered as a confounder and did not show a strong influence on premenopausal breast cancer is non-negligible and that this issue may result in some bias. As is usual practice in epidemiological studies, relative risks were provided for several categories of BMI. To overcome this problem the authors estimated a regression coefficient for the relative risk as a function of the BMI, however, several critical assumptions are necessary for this type of approach. The authors found severe heterogeneity across all studies combined (the p-value of a corresponding test was smaller than 10^{-8}). An influence of the type of study (cohort-study or case-control study) was apparent. Therefore, no overall summary is presented for case-control and cohort studies combined. However, the authors present a summary estimate for all case-control studies, although the severe heterogeneity ($p < 10^{-8}$) was still present. One reason for the heterogeneity is the difference in adjustment for confounders. Adjustment for confounders other than age was used only in 10 out of the 23 studies. Several other issues may have caused the severe heterogeneity between studies and the summary assessment of an inverse association of high BMI with risk of premenopausal breast cancer must be interpreted with caution.

White (1999) investigates the level of alcohol consumption at which all-cause mortality is least. Based on a MEDLINE search he included 20 studies in the meta-analysis; nine studies were excluded because information needed for the meta-analysis was not available. The heterogeneity of the study populations and the differences of confounding factors is obvious from the summary table in the paper. Age is the only factor used in all studies, the number of additional factors ranges from 0 to 12. The author has used a complex method to re-calculate unadjusted relative risk estimates, however, combining crude estimates may yield a severe bias if confounding plays a role. Additionally, various numbers of categories based on different cutpoints were used in the individual studies. To combine these different estimates, the author fit asymmetric quadratic functions to the results based on categorized alcohol levels in the original papers. From each function a nadir indicating the least all-cause mortality was estimated. Obviously, the nadir depends strongly on the original chosen categories and the calculation to estimate the nadir from the published data are questionable and could yield a major bias. Four studies that did not show a "typical" U-shape relationship were considered to be noninformative about the nadir and were excluded from the analysis. For different subpopulations estimates of the nadir with corresponding confidence intervals were presented. The estimate is much higher for UK men than for US men, however, because of many limitations of the MAL the results are questionable (Sauerbrei, Blettner, & Royston, 2001). The used approach gives estimates that may be wrong and may lead to possibly wrong recommendations regarding the alcohol consumption. Using the publications only, a careful investigation of the large heterogeneity between studies and countries could have been a worthwhile exercise, but the calculation of a quantitative estimate is meaningless.

Based on published data Ben-David, Rosen, Franssen, Einarson, and Szyfer (1995) present a meta-analysis investigating the influence of dose intensity of first line chemotherapy with cis- or carboplatin alone or in combination with other chemotherapy drugs on median survival of stage III-IV ovarian cancer patients. Following some central rules for meta-analyses they identified 61 "separate units", some from randomized trials and some from observational studies, which can be seen as independent one-armed observational studies. Based on the intended dose for each study and the distribution of the prognostic factors given in the corresponding publication, they tried to identify the influence of dose intensity of platin (DI), total dose intensity of platin (TDI), and total dose intensity of all chemotherapy drugs (GDI). The amount of information about prognostic factors given in the published papers and their incorporation in adjusting the treatment effect varied substantially between the different studies. As the effect of prognostic factors on survival is much stronger than the benefit from the treatment, observational studies require their incorporation in a multivariate model for the estimation of a treatment effect. However, in the published papers this issue is handled differently, leading to estimates which cannot be combined in a sensible way. Median survival categorized as less than 20 versus more than 20 months was used as the only outcome for each study. This very simple measure is not informative as it does not use survival time per patient. This choice was certainly guided by simplicity because only published papers were used. In a comment on the paper, Sauerbrei, Blettner, and Schumacher (1996, p. 428) concluded that

> in contrast to the obvious effect of TDI on median survival presented in the paper we believe that the authors did not succeed in adding any important information to the question of dose intensity on the survival of ovarian cancer patients.

Furthermore, a large randomized trial convincingly reached a result in contrast to that of the meta-analysis presented by Ben-David et al. (1995). Hopefully, not too many clinicians read this oversimplified meta-analysis and came to wrong conclusions for the treatment of their patients.

In breast cancer more than 100 factors are discussed as being potentially prognostic, however, only the number of positive axillary lymph nodes is generally accepted to have an important influence on prognosis. For most factors, only unsystematically traditional reviews have been published. The confusion caused by the current way of summarizing the evidence will be demonstrated by citing some review papers on HER2/neu oncogene, also known as c-erbB-2, a factor of strong interest in the last years. For this factor several traditional reviews based on different included studies, using different methods and leading to different conclusions have been published.

Based on their review, Allred, Harvey, Berardo, and Clark (1998) conclude that HER2/neu is at best a weak prognostic factor in node negative patients, whereas it seems to have a more pronounced prognostic value in node positive patients. More detailed information of single studies is given by Révillion, Bonneterre, and Peyrat (1998). They state that in several studies the prognostic value of HER2/neu was present only in univariate analysis, and not in multivariate analysis. Furthermore, different results are sometimes reported for the effect on disease-free survival and overall survival. The tables 7 to 9 by Révillion et al. (1998) give evidence that incorporation of classical factors varies severely between the different studies and that, despite of the strong relationship to estrogen and progesterone receptors, these factors are often not considered in the analysis. For different studies this may cause severe differences in the estimated effect of the prognostic value of HER2/neu and makes a sensible comparison of results almost impossible. Révillion et al. (1998) summarize: "In univariate analyses HER2/neu is strongly associated with poor prognosis. However, it does not retain a clinical prognostic significance in multivariate analyses since it is associated with several strong prognostic parameters" (p. 791). As this statement is only based on a listing of the classical prognostic parameters used in the individual studies and on the simple assessment of significance for each single study but without any discussion on the power in the "negative" studies or on the size of the effect if the factor had a significant influence we consider the scientific basis for these important statements as being very low. Their review clearly demonstrates the heterogeneity between the studies concerning treatment, follow-up time, and on the issue of subgroup analyses. From our point of view any summary assessment seems unjustifiable. In another review published in two papers Ross and Fletcher (1998, 1999) list 47 studies investigating the prognostic value. They do not care about mixing results from univariate and multivariate analyses of the single studies, whose results are simply given as impact on prognosis "yes" or "no". In the paper published first they conclude "The preponderance of evidence indicates that HER2/neu gene amplification and protein overexpression are associated with an adverse outcome in breast cancer" (Ross & Fletcher, 1998, p. 424). A reader gets certainly the impression of an important prognostic factor. In the latter paper they give no summary statement but concentrate more on different measurement techniques and list studies with significant and non-significant results, respectively. Obviously, a useful summary assessment of the prognostic value seems impossible with the traditional review. In their later paper they seem to have realized it, however, they did not explicitly state it. From our point of view such a statement would have added some value to their paper.

Difficulties in general associated with reviews of prognostic factor studies are discussed in Altman and Lyman (1998). Most problems are closely connected to issues of assessing the importance of a factor from the results of observational studies.

6.7 CONCLUSION

This chapter has described and critically assessed different review methods for observational studies. We strongly believe that all available data and information are needed for full assessment of weak factors and that systematic reviews of available evidence will become increasingly important. A major impediment for meta-analysis of observational studies is the heterogeneity between studies in their design, data collection methods, and statistical modeling. Mainly because of the last aspect meta-analyses using published data are, therefore, limited and give rarely a valid quantitative estimate or dose-response function. However, a meta-analysis of published data may be more reproducible than a qualitative review. A MAL has the trivial but dangerous advantage of being less expensive and time consuming than a meta-analysis with individual data. Consequently, some authors will continue to publish results from those meta-analyses and public health regulators, and decision-makers may rely on these results, even if the scientific value is questionable. Therefore, it remains important to point to the weaknesses and flaws in meta-analyses of literature. In particular, errors and bias that can be produced when combining studies with different design, methods, and analytic models need to be addressed. Despite of the large costs in time and manpower researchers should be encouraged to aim for meta-analyses with patient data. Several successful projects have shown that it is possible to interest researches all over the world for the collaboration (Advanced ovarian cancer trialists group, 1991; Early Breast Cancer Trialists' Collaborative Group, 1992), mainly because the question was so important that the scientific community was strongly interested in scientific answers. We believe that a useful MAP does not always need to incorporate all studies conducted for the specific question of interest but that a well defined "group of studies" – for example, only new, good, and large studies – may be sufficient. Such an approach will substantially reduce the costs and largely increase the probability to receive the individual data from the studies of interest. A meta-analysis starting with a re-analysis of the individual studies would have a chance to result in valid estimates or dose-response functions. With our examples we tried to show that the more traditional ways have often failed to give a reliable assessment for a factor of interest, despite of the fact that an enormous amount of money was spent from the individual study groups all over the world and data are available on tens of thousands of patients.

Statistical methods for pooling data from different sources have to be refined and new approaches are needed. Some important work is currently in progress, for example, from members of the "Statistical Methods Working Group" of the Cochrane Collaboration. Methods for conducting and reporting of meta-analyses of published data need to consider the basic limitations. While significant progress has been made in the systematic approaches for meta-analyses of randomized clinical trials, limitations in observational studies may not be overcome by too simple statistical methods. However, an equally rigorous standard is needed as more public health decisions will be relying on the results of meta-analyses. Hence, the research community must ensure that the validity, reliability, and overall quality of these methods is improved.

REFERENCES

Advanced ovarian cancer trialists group. (1991). Chemotherapy in advanced ovarian cancer: An overview of randomized clinical trials. *British Medical Journal*, *303*, 884–893.

Allred, D. G., Harvey, J. M., Berardo, M., & Clark, G. M. (1998). Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern Pathology*, *11*, 155–168.

Altman, D. G., & Lyman, G. H. (1998). Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, *52*, 289–303.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts – Treatment for myocardial infarction. *Journal of the American Medical Association*, *268*, 240–248.

Bailar III, J. C. (1997). The promise and problems of meta-analysis. *New English Journal of Medicine*, *337*, 559–561.

Ben-David, Y., Rosen, B., Franssen, E., Einarson, T., & Szyfer, I. (1995). Meta-analyses comparing cisplatin total dose intensity and survival. *Gynecolocic Oncology*, *59*, 93–101.

Blettner, M., & Sauerbrei, W. (1993). Influence of model-building strategies on the results of a case-control study. *Statistics in Medicine*, *12*, 1325–1338.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., & Friedenreich, C. M. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*, *28*, 1–9.

Boffetta, P., Saracci, R., Andersen, A., Bertazzi, P. A., Chang-Claude, J., Cherrie, J., Ferro, G., Frentzel–Beyme, R., Hansen, J., Plato, N., Teppo, L., Westerholm, P., Winter, P. D., & Zochetti, C. (1997). Cancer mortality among man-made vitreous fiber production workers. *Epidemiology*, *8*, 259–268.

Chalmers, T. C. (1991). Problems induced by meta-analyses. *Statistics in Medicine*, *10*, 971–980.

Chalmers, T. C., & Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research*, *2*, 161–172.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurence. *Journal of the American Medical Association*, *263*, 1385–1389.

Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *Aids Education and Prevention*, *9 (Supplement A)*, 15–21.

Dickersin, K., Scherer, R., & Lefebre, C. (1994). Identifying relevant studies for systematic reviews. *British Medical Journal*, *309*, 1286–1291.

DuMouchel, W. (1990). Bayesian metaanalysis. In D. A. Berry (Ed.), *Statistical methodology in the pharmaceutical sciences* (pp. 509–529). New York: Dekker.

Early Breast Cancer Trialists' Collaborative Group. (1992). Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomized trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *The Lancet*, *339*, 1–15.

Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal*, *309*, 789–792.

Friedenreich, C. M. (1993). Methods for pooled analyses of epidemiologic studies. *Epidemiology*, *4*, 295–302.

Galbraith, R. F. (1994). Some applications of radial plots. *Journal of the American Statistical Association*, *89*, 1232–1242.

Gilbert, E. S. (1989). Analyses of combined mortality data on workers at the Hanford Site, Oak Ridge National Laboratory and Rocky Flats Nuclear Weapons Plant. *Radiation Research*, *120*, 19–35.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351–379.

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiological Review*, *9*, 1–30.

Greenland, S. (1994). Invited commentary: A critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, *140*, 290–296.

Lubin, J. H., Boice, J. D. J., Edling, C., Hornung, R. W., Howe, G., Kunz, E., Kusiak, R. A., Morrison, H. I., Radford, E. P., Samet, J. M., & et al. (1995). Radon-exposed underground miners and inverse dose-rate (protraction enhancement) effects. *Health Physics*, *69*, 494–500.

Mengersen, K. L., Tweedie, R. L., & Biggerstaff, B. J. (1995). The impact of method choice on meta-analysis. *Australian Journal of Statistics*, *37*, 19–44.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining and reporting. *Statistics in Medicine*, *18*, 321–359.

Paul, S. R., & Donner, A. (1989). A comparison of tests of homogeneity of odds ratios in K 2 × 2 tables. *Statistics in Medicine*, *8*, 1455–1468.

Petitti, D. B. (1994). *Meta-analysis, decision-analysis and cost-effectiveness analysis. Methods for quantitative synthesis in medicine.* Oxford: Oxford University Press.

Révillion, F., Bonneterre, J., & Peyrat, J. P. (1998). ERBB2 oncogene in human breast cancer and its clinical significance. *European Journal of Cancer*, *34*, 791–808.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Ross, J. S., & Fletcher, J. A. (1998). The HER–2/*neu* oncogene in breast cancer: Prognostic factor, predictive factor and target for therapy. *Stem Cells*, *16*, 413–428.

Ross, J. S., & Fletcher, J. A. (1999). The HER–2/*neu* oncogene: Prognostic factor, predictive factor and target for therapy. *Cancer Biology*, *9*, 125–138.

Sauerbrei, W., Blettner, M., & Royston, P. (2001). On alcohol consumption and all-cause mortality, letter to White. *Journal of Clinical Epidemiology*, *54*, 537–538.

Sauerbrei, W., Blettner, M., & Schumacher, M. (1996). Re: Meta-analysis comparing cisplatin total dose intensity and survival: A critical reappraisal. *Gynecologic Oncology*, *62*, 427–428.

Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. *American Journal of Epidemiology*, *140*, 771–778.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, *14*, 2685–2699.

Spiegelhalter, D. J., Thomas, A., Best, N., & Gilks, W. (1996). BUGS: Bayesian inference Using Gibbs Sampling. Available from http://www.mrc-bsu.cam.ac.uk/bugs/.

Stewart, L. A., & Clarke, M. J. (1995). On behalf of the Cochrane working group on meta-analysis using individual patient data. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, *14*, 2057–2079.

Stewart, L. A., & Parmar, M. K. B. (1993). Meta-analysis of the literature or of individual patient data: Is there a difference? *The Lancet*, *341*, 418–422.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal*, *309*, 1351–1355.

Thompson, S. G., & Pocock, S. J. (1991). Can meta-analyses be trusted? *The Lancet*, *338*, 1127–1130.

Ursin, G., Longenecker, M. P., Haile, R. W., & Greenland, S. (1995). A meta-analysis of body mass index and risk of premenopausal breast cancer. *Epidemiology*, *6*, 137–141.

White, I. R. (1999). The level of alcohol consumption at which all-cause mortality is least. *Journal of Clinical Epidemiology*, *52*, 967–975.

Ziegler, S., & Victor, N. (1999). Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität [Some problems of the standard meta-analysis methods in the presence of heterogeneity]. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, *30*, 131–140.

7

Meta-Analysis of Randomized Clinical Trials in the Evaluation of Medical Treatments – A Partly Regulatory Perspective

Armin Koch
Joachim Röhmel

Federal Institute for Drugs and Medical Devices
Berlin

Summary

Meta-analysis of randomized clinical trials plays an important role in summarizing available evidence with respect to the comparison of different drugs for the same indication. In contrast, up to now meta-analysis is of only minor importance in the process of new drug application despite the fact that also in this situation a summary evaluation of available evidence from a, although limited, number of independent clinical trials is necessary. The main reason is, in our opinion, that presented meta-analyses often are not completely convincing because objectives are not appropriately chosen and conduct or presentation are not sufficiently detailed so that the reader can assess provided evidence. This chapter is intended to clarify why some meta-analyses have higher credibility than others and provide some guidance to how the credibility of meta-analyses can be increased. Presented ideas are, hopefully, not only in the regulatory setting of importance.

7.1 INTRODUCTION

Meta-analysis has been defined to be a quantitative and systematic summary of a collection of separate studies for the purpose of obtaining information that can not be derived from any of the studies alone (Boissel et al., 1988). With this definition, meta-analysis implicitly is also a technique that should lead to reproducible results and that can be distinguished from the classical review or overview, where results from various studies might be collected and qualitatively weighted by an expert in the field.

Originally invented in the social sciences, meta-analysis has found widespread use in clinical research during the last two decades and the per-year number of published meta-analysis is still increasing. However, only in rare cases has the discussion about the appropriateness of biostatistical methodology in medical research been as intensive as was the case with meta-analysis. From the very beginning meta-analysis has split up the community into clear proponents and those who completely dislike this type of analysis.

Feinstein (1995) named meta-analysis a synonym for "statistical alchemy for the 21st century", and others expressed their doubts on the credibility of results "proven" by means of meta-analysis. It has repeatedly been emphasized that pivotal trials should be designed to stand on their own and that in consequence meta-analyses should not be necessary ("If a treatment has an effect so recondite and obscure as to require meta-analysis to establish it I would not be happy to have it used on me" (Eysenck, 1994, p. 792)). And also empirical comparisons of the results from meta-analyses with results from large randomized clinical trials (Villar, Carroli, & Belizan, 1995) or critical expert reading of meta-analyses do not support the hypothesis that a meta-analysis can replace randomized clinical trials ("In my own review of selected meta-analyses, problems were so frequent and so serious, including bias on part of the meta-analyst, that it was difficult to trust in the overall 'best estimates' that the method often produces" (Bailar, 1997, p. 560)).

A positive view on meta-analysis is best summarized by a citation from a recent paper by Resch (1996), who wrote:

> I disagree, however, that a meta-analysis should exclusively be viewed as "hypothesis generating". This proposal denies the fact that, however biased, a high-quality meta-analysis quantitatively summarizes the existing evidence. What could be a better basis for a clinician's treatment decision at the time it must be made? (p. 621)

Meta-analyses – being retrospective and non-experimental investigations – are in a strict sense observational studies (Victor, 1995). Comparing, however, the evidence gained from a prospective observational study and a meta-analysis based on randomized clinical trials, clarifies that this can not be the whole truth, as the latter are based on what has often been termed "best available evidence", and at least on the study level distribution of covariates is controlled.

Approval of a new drug by national or European agencies is one of the most sensitive areas of evaluation of knowledge provided by clinical trials: If in a certain indication there is not yet a standard treatment, the approval of a new drug defines this standard and all future developments will be validated against this standard. In addition, an acceptable benefit/risk-ratio is in general taken for granted whenever a new drug is licensed.

International guidelines on statistical principles in clinical trials (ICH Topic E9) have made a valuable contribution to clarifying methodological principles for clinical trials in the regulatory setting. In this document reference is made to the use of meta-analysis or pooled analyses in general, and subsequently various meta-analyses have been presented in new drug applications. Not in all cases have these analyses been appropriate from a regulatory viewpoint, and the need for further clarification became obvious. This is also reflected in the current attempt to harmonize the opinions of the European regulatory authorities in a Points to Consider document, which is intended to provide better guidance for the pharmaceutical industry.

This chapter is not intended to summarize the discussion on the development of the Points to Consider document. Instead, ICH-E9 statements on meta-analysis are briefly reviewed and given recommendations are summarized. Instances are mentioned where these recommendations might need additional clarification or give the impression that the view on meta-analysis in the regulatory setting is very narrow. From a scientific viewpoint arguments why credibility of meta-analysis is in some instances greater than in others are given, and factors that can influence credibility are named. Some sample situations are discussed for illustration. A more appropriate use of meta-analysis will hopefully help the technique find greater acceptance in the regulatory setting.

7.2 QUOTES FROM "THE GUIDELINE"

In chapters II (Considerations for overall clinical development) and chapter VII (Reporting) direct or indirect reference is made to techniques of summarizing results from more than one clinical trial. The exact wordings are:

> Interpretation and assessment of the evidence from the total programme of trials involves synthesis of the evidence from the individual trials. This is facilitated by ensuring that common standards are adopted for a number of features of the trials such as dictionaries of medical terms, definition and timing of the main measurements, handling of protocol deviations and so on. A statistical summary, overview or meta-analysis may be informative when medical decisions are addressed in more than one trial. Where possible this should be envisaged in the plan so that the relevant trials are clearly identified and any necessary common features of their designs are specified in advance. Other major statistical issues (if any) that are expected to affect a number of trials in a common plan should be addressed in the plan. (Section 2.1.1: Development plan)

An overall summary and synthesis of the evidence on safety and efficacy from all the reported clinical trials is required for a marketing application [...]. This may be accompanied, when appropriate, by a statistical combination of results. [...] addressing the key questions of efficacy by considering the results of the relevant (usually controlled) trials and highlighting the degree to which they reinforce or contradict each other; [...]. During the design of a clinical programme careful attention should be paid to the uniform definition and collection of measurements which will facilitate subsequent interpretation of the series of trials, particularly if they are likely to be combined across trials. A common dictionary for recording the details of medication, medical history and adverse events should be selected and used. A common definition of the primary and secondary variables is nearly always worthwhile, and essential for meta-analysis. The manner of measuring key efficacy variables, the timing of assessments relative to randomization/entry, the handling of protocol violators and deviators and perhaps the definition of prognostic factors, should all be kept compatible unless there are valid reasons not to do so. Any statistical procedures used to combine data across trials should be described in detail. Attention should be paid to the possibility of bias associated with the selection of trials, to the homogeneity of their results, and to the proper modeling of the various sources of variation. The sensitivity of conclusions to the assumptions and selections made should be explored. (Section 7.2: Summarizing the clinical database)

Individual trials should always be large enough to satisfy their objectives. Additional valuable information may also be gained by summarizing a series of clinical trials which address essentially identical key efficacy questions. The main results of such a set of trials should be presented in an identical form to permit comparisons, usually in tables or graphs, which focus on estimates plus confidence limits. The use of meta-analytic techniques to combine these estimates is often a useful addition, because it allows a more precise overall estimate of the size of the treatment effects to be generated, and provides a complete and concise summary of the results of the trials. Under exceptional circumstances a meta analytic approach may also be the most appropriate way, or the only way, of providing sufficient overall evidence of efficacy via an overall hypothesis test. When used for this purpose the meta-analysis should have its own prospectively written protocol. (Section 7.2.1: Efficacy data)

In summarizing safety data it is important to examine the safety database thoroughly for any indications of potential toxicity, and to follow up any indications for an associated and supportive pattern of observations. The combination of the safety data from all human exposure to the drug provides an important source of information, because its larger sample size provides the best chance of detecting the rarer adverse events and, perhaps of estimating their approximate incidence. [...] The results from trials which use a common comparator (placebo or specific active comparator) should be combined and presented separately for each comparator providing sufficient data. (Section 7.2.2: Safety data)

In summary, ICH-E9 makes the following proposals and recommendations on the use of meta-analysis:

1. Objectives for meta-analyses might be the gain of more precise overall estimates of the size of the treatment effect, the gain of a complete and concise summary of trial results, and the assessment of consistency of results across trials.
2. Meta-analysis should be prospectively planned with the clinical trials programme in the development of a new treatment. This is extremely important if the meta-analysis is the most appropriate or the only way to provide sufficient evidence of efficacy.
3. Sensitivity analyses are necessary if assumptions or selections are necessary to justify the combination of study results.
4. Safety-results from trials where the same active treatment has been compared with different active substances or placebo should not be combined into an overall estimate of "a treatment effect". In general, separate analyses should be presented for different comparators.

Some other recommendations might be seen too narrow or even contradictory:

1. Studies should address essentially identical efficacy questions (not necessary: The only question is whether the design of the studies and the query for variables, relevant for the meta-analysis, is sufficiently similar)
2. Presentation of results of the single trials should be done by means of estimates and confidence intervals (better: In general, the number of successes and events per treatment group for dichotomous endpoints or sufficient statistics in general should be provided in order to give the reader the option to make his own mind).
3. "Studies should stand on their own" and "studies should address essentially identical key efficacy questions" (little dissent exists on how to interpret the results of meta-analysis in this situation).

And in some instances further clarification is needed:

1. No guidance is given with regard to the exceptional circumstances, for which a proof of efficacy by means of a meta-analysis might be acceptable.
2. Supposed it is accepted that a meta-analysis can increase the precision of an estimate for the treatment effect in a certain situation, then meta-analysis can in fact be confirmatory.
3. The term "prospectively" needs clarification: Should the meta-analysis be planned before the first study that is intended to be included in the meta-analysis is undertaken, or is it sufficient to plan the meta-analysis before its conduct.

7.3 HOW CAN THE CREDIBILITY OF META-ANALYSIS BE INCREASED?

Meta-analyses are non-experimental studies. As a consequence, the classical framework of error probabilities is not applicable for decisions based on p-values or confidence intervals that have been computed in a meta-analysis of results from independent trials. Like in observational studies in general, p-values are primarily a measure for the distance between two success-rates or between two means computed in two different groups in relation to the respective variance. And like in observational studies, too, the author of a meta-analysis must justify his belief that observed differences between two groups, defined by means of the absence or presence of a certain treatment, are in fact due to differences between the two treatments and not a consequence of bias (e.g., due to selection of trials (i.e., publication bias), patients and interventions, or the statistical methodology for the combination of the results from independent trials).

Results and conclusions from meta-analysis are thus more or less credible and this credibility – in contrast to randomized clinical trials – does not only depend on design issues but also on sound argumentation on the absence of bias. The following sections name the factors that influence to our opinion this credibility and name prerequisites for meta-analysis, performed "in an almost confirmatory way". A number of situations are proposed where meta-analysis therefore can contribute valuable information for the decision on licensing of new drugs.

7.3.1 The Aspect of Objectives for Meta-Analyses

For a long time, meta-analyses have had little impact on decisions made by regulatory authorities. This was mainly due to the fact that meta-analyses have been presented almost exclusively in situations where proof of efficacy in two independent clinical trials deemed necessary in the beginning, and this attempt failed in the end (i.e., one significant and one insignificant trial; two only borderline significant results etc.). Whenever meta-analyses are misused to counterbalance for shortcomings in the respective primary research (i.e., the studies to be included in the meta-analysis), credibility is affected: The only aim of this type of meta-analysis is the demonstration of a "significant" treatment effect. Chances are not too bad that at least this aim is reached due to the mere fact that sample size is increased. This unpleasant situation has postponed considerations on good objectives for meta-analyses in the regulatory setting:

1. Substantiation of additional claims on secondary endpoints, especially in a situation where the primary endpoint is based on a surrogate-variable, or on components of multiple endpoints:

   Various situations exist, where in a collection of clinically important variables the choice of the primary endpoint is not only driven by the attempt to select the most important one, but also by feasibility considerations. This is especially true if the incidence of some of the endpoints is higher and some other endpoints are rare. An example is provided by studies investigating postoperative prophylaxis against thromboembolic complications, where the incidence of deep vein thromboses is usually selected as the primary endpoint. The rare event of a pulmonary embolism is, however, an at least equally important endpoint. The demonstration of equivalence or superiority or equivalence with respect to the rate of the rare event would demand larger clinical trials. A good basis for a claim with respect to the rare endpoint might be a meta-analysis of all pivotal trials.

Similarly, a double endpoint (death or reinfarction) is selected in studies in the treatment of acute myocardial infarction. An additional claim "the experimental treatment reduces the rate of death" may be substantiated by means of a meta-analysis of all pivotal clinical trials.

2. Proof and investigation of efficacy and safety in subgroups of the patient population:

   As soon as the global superiority of an experimental treatment over control has been established by means of evidence from more than one pivotal clinical trial, investigations into subgroups of the patient population might help to understand better for which patients the benefit of the experimental treatment is greatest (e.g., to justify higher costs of the experimental treatment). Similarly, one might be conscious whether a consistent safety profile exists across subpopulations. In both instances a meta-analysis can be helpful to provide evidence based on an acceptable sample size.

3. Proof of efficacy in situations where single studies are contradictory or inconclusive:

   Despite the fact that meta-analysis should not be used to compensate for shortcomings in the pivotal trials, meta-analysis can be helpful to come to a decision if results of clinical trials are not homogeneous. In this situation meta-analysis must be understood as a tool to demonstrate robustness of results and conclusions by means of sensitivity analyses. In general, the same methods should be applied that were proposed for the investigation of heterogeneity in multicenter clinical trials. It should be noted that, due to currently existing limitations with respect to statistical methodology, two studies can not be regarded sufficient to assess an overall "impression" by means of a meta-analysis. This is due to the fact that the likelihood to detect important differences between studies with respect to the treatment effect is small.

4. Evaluation of signals for serious adverse events of treatment:

   In many situations a meta-analysis of "all randomized evidence" will be the only chance to detect early whether a risk for serious adverse events is associated with the experimental treatment at the time a decision on marketing of the new drug has to be made. Whereas in early years simply a summary of the combined database was provided, during the last years it has been recognized that formal methods for the combination of results from independent studies should also be used in this situation to avoid bias.

7.3.2 The Aspect of Planning

Planning of experimental investigations in humans or animals is mandatory by law. As has been pointed out in the discussion on the need for randomization in clinical trials, the object of trials is both to ensure a high probability of identifying the better treatment (if there is one) and to convince others of the validity of the conclusions (Byar et al., 1976).

In consequence, all scientific investigations and especially observational research should be planned. The crucial point in the context of meta-analysis is that the need for a meta-analysis can become obvious at various points during the conduct of a clinical programme consisting of more than one clinical trial.

It is, however, not only a question of credibility of results whether such a meta-analysis has been planned together with the whole clinical programme or after completion of the most recent study in the program: Credibility is affected if the applicant can not assure that presented conclusions are not driven by observed results and ruling this out is obviously easier in the first case.

One might believe that the credibility of a meta-analysis planned after the completion of the last clinical trial might be increased if the meta-analysis has been performed by a site that is independent or quasi-independent from the sponsor. Even in this situation it might be difficult to assure that results of the meta-analysis have not been known to the sponsor before the "independent" re-analysis has been performed.

When planned in the beginning of the clinical program, the additional opportunity exists to care for consistency in the conduct of the clinical trials that are intended to be included into the meta-analysis. This reduces the need to make assumptions on what can be safely combined (e.g., a study lasting three weeks and another study lasting four weeks, studies where variables have been transformed differently in different studies, or studies with slightly different questionnaires that might affect clinicians behavior with respect to answering), reduces potential sources of heterogeneity, and thus also improves the quality of the meta-analysis.

A meta-analysis planned before the inception of the last study in a clinical program ranges in between the before mentioned extremes: The conformity of the study plans can at this point hardly be influenced. However, which endpoints and which analysis are presented in the end can not be completely derived from the observed results but at least some sort of internal validation is available. This is the reason why one positive meta-analysis and a subsequently initiated clinical study with positive results may constitute a sufficient basis for a licence application (see Example 2 in Section 7.4).

A minimal requirement for meta-analysis in the regulatory setting is that the meta-analysis has been planned in advance to its conduct. In this situation it is, of course, difficult to demonstrate that results have not been available at the point in time, where the plan for the meta-analysis has been presented.

7.3.3 The Aspect of Conduct

Due to unfavorable experiences with publication based meta-analysis recently, meta-analyses based on individual patient data have been recommended and termed the current gold standard in meta-analysis (Clarke & Stewart, 1994). To our present opinion and experiences this might be somewhat too restrictive. Obviously, more questions can be addressed in a meta-analysis based on individual patient data, as the full information on covariables is available. It is also true that more insight into the data at hand is needed to perform this type of meta-analysis. Keeping in mind, however, the enormous workload that is needed to perform a re-analysis of the individual trials, one should also keep in mind that the method which is used for combination of study results should be justified by the question that is to be answered (e.g., in case subgroup analyses are of interest and the respective information is not available from the study report, a re-analysis can not be avoided).

At least in the regulatory framework and with respect to the primary and secondary end-points of pivotal trials, in contrast, it would shade suspicion on the original trial report if results of meta-analyses based on individual patient data and meta-analyses based on published data from the original report would come to different conclusions. Re-definitions of success and treatment failure, in addition, might raise suspicion that again attempts are made to fish for significance (i.e., why should definitions that seemed reasonable at the time when the individual trials had been planned now be obsolete?). Obviously, the plan to perform a meta-analysis based on original patient data can not be the justification for the exclusion of trials from the analysis where original patient data are not available, although it is expected that this problem (like publication bias more generally) is of minor importance in the regulatory setting.

7.3.4 The Aspect of Analysis and Presentation of Results

During the first years the term meta-analysis has been associated in medicine almost completely with the aspect of summarizing the evidence from independent clinical trials. Summary estimates and overall tests of effect or confidence intervals have been presented exclusively. Due to bad experiences the view on meta-analysis is nowadays more differentiated and meta-analysis is more understood as a tool for investigating similarities and dissimilarities between trials that should, at least in principle, be combinable. Unfortunately, in the regulatory setting the mere provision of a summary p-value is still the rule and not the exception. As the evaluation of consistency of results across trials is very important for claims on efficacy, this is not acceptable.

Statistical information on similarities and dissimilarities of study results are important. Critics might say that the currently used tests for homogeneity that are based on weighted squared differences between estimates from the single trials and the meta-analysis estimate have insufficient power to detect departures from the null-hypothesis (Jones, O'Gorman, Lemke, & Woolson, 1989). Critics might further object that they are in addition usually used "to proof their null-hypothesis". As a consequence, it should be not acceptable to conclude that no heterogeneity exists, unless the test for homogeneity rejects the null-hypothesis at a conventional 5%-level. This is true but should, to our opinion, not prevent from making all attempts to use the test as a diagnostic tool (as a well known statistician has pointed out: Statistical methods need not be perfect, it is sufficient if they are better). The following example might support this opinion:

Two randomized double blind placebo controlled studies have been undertaken to investigate the efficacy of omeprazole in functional dyspepsia (Bond study and Opera study) (Talley et al., 1998). Both studies are three armstudies comparing two dosage regimens to placebo. Results of the comparison of the higher dose and placebo are reported here in a slightly simplified way not to discredit a potentially efficacious treatment but to construct an instructive example for decision making. Conclusions of the authors are: Omeprazole is modestly superior to placebo in functional dyspepsia. On an intention to treat analysis ($n = 1248$), complete symptom relief was observed in 38% on omeprazole 20mg compared to 28% on placebo ($p = .002$).

Table 7.1 Complete Relief of Dyspeptic Symptoms

Study	Omeprazole 20mg Relief / Treated	Placebo Relief / Treated
Bond	93 / 219	57 / 219
Opera	68 / 202	62 / 203

Results of the two studies are summarized in Table 7.1. The paper reports results for the combination of the two trials only. The meta-analysis estimate for the difference of the relief rates in the two treatment groups is 10% with a 95% confidence interval ranging from 3.7 to 16.3%, and from this a modest superiority of the experimental treatment over placebo is concluded.

Our confidence in the drawn conclusion might change if it was clearly stated, that first, it is only the Bond-study that came up with a significant treatment effect ($p = .001$), the p-value for the treatment effect in the Opera study was $p = .501$, and that second, despite the fact that tests for heterogeneity are blamed for being insufficient, a clear warning might have been achieved ($p = .039$ for heterogeneity).

This should be general guidance for analysis and presentation of results: Confinement to only meta-analytic results in terms of a summary estimate of the treatment effect and the respective confidence interval or a simple hypothesis test resulting in just one p-value is not appropriate in the setting of observational studies. As has been pointed out before, argumentation is necessary with observational studies. Analysis and presentation of results should always emphasize the need to also clarify the contribution of a single trial to the combined result. Respective recommendations date back to 1993 (Thompson, 1993): For every study the relative weight in a fixed effects model should be presented together with the contribution of the study to the statistics of the heterogeneity test. The first information gives the reader an impression on whether meta-analysis can add useful information to the knowledge from the larger studies (e.g., in a situation with one large trial and two small studies given weights 80%, 10%, 10%, it is very unlikely that the combined analysis will add new information, as the size of the estimate is completely driven by the result of the large trial). The second information can descriptively be assessed with a rule of thumb, comparing each of the contributions to the heterogeneity statistics with the critical value of a χ^2-distribution with one degree of freedom and deciding whether results are homogeneous or whether some extreme results might drive the overall impression.

7.4 SAMPLE SITUATIONS

A series of sample situations of appropriate or inappropriate use of meta-analysis, all motivated by recent applications or collected from the literature, are presented to illustrate the considerations above.

Example 1: Meta-Analysis and Borderline Significant Pivotal Studies

In a situation where proof of superiority has to be based on two separate pivotal trials, both demonstrating that the experimental treatment is superior to control, both studies ended up with only borderline significant results (e.g., a p-value between 5% and 10% was achieved, where the level of significance was initially set to 5%). A combined re-analysis of the two pivotal trials demonstrates "significant" superiority of the experimental treatment over control for the primary endpoint of the pivotal trials.

Even if the meta-analysis has been planned before its conduct, a meta-analysis in this situation is not acceptable. Meta-analysis should not be used as a safety-belt against non-significant results from pivotal trials that were planned to stand on their own.

In a very dialectic discussion on meta-analysis Senn (1997) argued that clinical trials are notoriously too small and that even small true effects might be of enormous public health importance (for example, in the treatment of cancer or myocardial infarction) and that many small drops can make a hole into stone. Nevertheless, he admits that a difference to the drug development setting exists, where the way how experimentation is performed is under the control of the pharmaceutical company that must demonstrate efficacy beyond reasonable doubt.

If studies that have been planned to stand on their own fail to proof efficacy as expected, this is a strong indication that something substantial went wrong with the original plan.

Example 2: Meta-Analysis and the Need for Replication of Results

Imagine a situation where the sponsor has decided to proof efficacy in a series of three phase III pivotal clinical trials. After a first "significant" trial the sponsor decides to make some minor modifications to the study design (e.g., small changes in the criteria for inclusion or exclusion of patients from the trial or a modification of the primary variable). His intention is to demonstrate even better the superiority of the experimental treatment over control.

Unfortunately, this second and a third attempt with again minor modifications both fail to demonstrate superiority of the experimental treatment over control. However, a meta-analysis including the first trial and "similarly defined subgroups" of the following two trials demonstrates a significant superiority of the experimental treatment over control. Again this is, in our opinion, no appropriate use of meta-analysis. If a need for replication of scientific results exists, this need can not be substituted by a (retrospective) subgroup (or a combination of subgroups) analysis. The sponsor still needs to demonstrate in an independent study that he now can correctly identify those patients that will benefit more from the experimental treatment than from the control treatment.

In consequence, the reversed situation might well be acceptable: Based on two (or more) non-significant trials the sponsor now believes that he can identify the patient population that will benefit from the experimental treatment. A meta-analysis of the respective subgroups of the first two trial populations demonstrates "significant" superiority of the experimental over the control treatment. A new trial is planned according to these restrictions and can verify the result of the meta-analysis. Meta-analysis should thus only be used to replace the first experiment, not the verification step.

The need for independent verification of scientific results has been discussed controversially in the literature (Högel & Gaus, 1999), and it has been even questioned whether the usual procedure of just performing two trials at the same time in different geographical hemispheres reflects a true verification of results. It should be pointed out that in this discussion the question is not whether verification is necessary or not, but that discrepancies between expectation and results need further investigation.

It should be noted that this is also one solution for the problem posed in Example 1: The meta-analysis of the two borderline significant results might, given that no other problems with study design and conduct exist, be accepted as a first pivotal trial. A third study should, however, be planned that can then successfully reproduce this first result.

Example 3: Meta-Analysis and Claims for Secondary Endpoints

Very often in clinical trials the selection of one variable as primary endpoint from a series of others, which then are termed secondary, is not only influenced by clinical importance of the various variables but reflects also considerations on feasibility (i.e., if a more important endpoint (e.g., pulmonary embolism in thrombosis prophylaxis) is a very rare event, chances for demonstrating superiority increase if an endpoint with higher incidence (e.g., deep vein thrombosis) is selected instead). A whole clinical trials program, however, might be designed such that also differences in mortality can be detected. The sponsor might decide to use the more frequent endpoint as primary, however, to plan a meta-analysis of pivotal trials in order to demonstrate superiority with respect to the less frequent event.

Given appropriate results, this is an acceptable prerequisite for an additional claim regarding the secondary, less frequent but potentially clinically more relevant endpoint.

7.5 CONCLUSIONS

Meta-analysis, even if restricted to the combination of only two pivotal trials, might be an extremely helpful tool for decision making in the regulatory setting. This technique could in principle support the task of the medical reviewer, who, at the end of the day, must integrate all the presented knowledge and come to a final decision. Meta-analysis is not playing this role up to now. This is mainly due to the fact that presented meta-analyses fall short with respect to the addressed objective, the conduct and the presentation of results: Still too much emphasis is given to combined estimates of the treatment effect and summary p-values. The potential of meta-analysis to show similarities and dissimilarities between the trials that are to be combined has not been used too often.

Meta-analysis, sometimes routinely presented as part of the clinical expert report, often fall short with respect to the presentation of results: Again, only summary information is presented and the reviewer is referred to the single documentation of clinical trials if he is interested in the consistency with respect to certain information.

Meta-analysis "is here to stay", however, careful consideration of the above mentioned points will help to sharpen the general understanding, where meta-analysis can be helpful and where not, will help to bring up better results and lastly help to find the place for meta-analysis that it should have.

REFERENCES

Bailar III, J. C. (1997). Editorial: The promise and problems of meta-analysis. *New England Journal of Medicine*, *337*, 559–561.

Boissel, J. P., Sacks, H. S., Leizorovicz, A., Blanchard, J., Panak, E., & Peyrieux, J. C. (1988). Meta-analysis of clinical trials: Summary of an international conference. *European Journal of Clinical Pharmacology*, *34*, 535–538.

Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselmann, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., & Ware, J. H. (1976). Randomized clinical trials. *New England Journal of Medicine*, *295*, 74–80.

Clarke, M. J., & Stewart, L. A. (1994). Obtaining data from randomized controlled trails: How much do we need for reliable and informative meta-analysis? *British Medical Journal*, *309*, 1007–1010.

Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal*, *309*, 789–792.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*, 71–79.

Högel, J., & Gaus, W. (1999). The procedure of new drug application and the philosophy of critical rationalism or the limits of quality assurance with good clinical practice. *Controlled Clinical Trials*, *20*, 511–518.

Jones, M. P., O'Gorman, T. W., Lemke, J. H., & Woolson, R. F. (1989). A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics*, *45*, 171–181.

Resch, K. L. (1996). Letter: Megatrials for clinical decision making. *Annals of Internal Medicine*, *125*, 621–622.

Senn, S. (1997). *Statistical issues in drug development.* New York: Wiley.

Talley, N. J., Meineche-Schmidt, V., Pare, P., Duckworth, M., Raisanen, P., Pap, A., Kordecki, H., & Schmid, U. (1998). Efficacy of omeprazole in functional dyspepsia: Double-blind, randomized, placebo-controlled trials (the Bond and Opera studies). *Alimentary Pharmacology and Therapy*, *12*, 1055–1065.

Thompson, S. G. (1993). Controversies in meta-analysis: The case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research*, *2*, 173–192.

Victor, N. (1995). "The challenge of meta-analysis": Discussion. Indications and contra-indications for meta-analysis. *Journal of Clinical Epidemiology*, *48*, 5–8.

Villar, J., Carroli, G., & Belizan, J. M. (1995). Predictive ability of meta-analysis of randomized controlled trials. *The Lancet*, *345*, 772–776.

8

Will it Work in Münster? Meta-Analysis and the Empirical Generalization of Causal Relationships

Georg E. Matt

Department of Psychology
San Diego State University

Summary

If vigorous physical exercise increases cognitive skills among elderly residents of San Diego, California, will it do the same for the elderly in Münster, Westphalia? This chapter examines the role of meta-analysis in justifying generalized causal inferences. In the *experimental tradition* of the social, behavioral, and natural sciences, such causal generalizations can be justified through a complete understanding of the causal conditions and mechanisms that bring about a phenomenon. Thus, rigorous experimentation and causal modeling of micro-mediating processes should provide the keys to valid causal generalization. In the *observational and correlational tradition* of the social, behavioral, and natural sciences, these generalizations are often justified through the correspondence between samples (or cases, instances, exemplars) and the populations (or universes, constructs, categories, classes) they are meant to represent. This emphasis on correspondence between samples and populations about which inferences are sought suggests that causal generalization may be best accomplished through rigorous random sampling. This chapter argues that, causal explanation and random sampling are of limited use for justifying generalized causal inferences because the causal moderating and mediating mechanisms are often poorly or incompletely understood and because random sampling – if at all possible – is infrequently practiced. Following a review of different validity models, Cook's (1990, 1993) five principles for strengthening empirical generalizations are presented in detail and illustrated in the context of meta-analysis. Finally, some conditions are outlined that promote generalizable inferences. The chapter concludes that valid empirical generalizations are best achieved through the synthesis of multiple studies, conducted by many research teams, with different populations, in different settings, with multiple operationalizations of interventions and outcomes. One form of research synthesis, meta-analysis, has particularly great promise to facilitate generalized inferences. Even though the best meta-analysis presents no shortcuts or guarantees for valid generalizations, it does provide research design and analytical tools to conduct principled investigations of generalizability claims, thus yielding stronger generalized inferences than are possible based on a single study alone.

8.1 INTRODUCTION

Imagine you just submitted a study for publication of the effects of physical exercise on cognitive skills in the elderly. The journal editor replies, wondering whether your findings apply only to the 49 volunteers from the retirement home across the street, your specific exercise regimen (i.e., ballroom dancing), the poorly-ventilated activity room at the retirement home, the specific operationalization of "cognitive skills" (e.g., standardized test involving verbal and numeric problems), and the foggy and cold November of 1999 when data were collected. Few scientific journals are interested in publishing a study if the research findings only apply to the unique circumstances in which they were conducted. In fact, textbooks on scientific methods (e.g., Babbie, 1995; Kerlinger, 1986) identify the pursuit of *general* truths as a defining feature of science. Do the findings from this study apply to other circumstances, possibly volunteers from other local retirement homes, other forms of exercise, better ventilated rooms, and different seasons? How can we justify such conclusions in the absence of strong sampling designs and strong causal explanatory theories?

This chapter deals with the empirical generalizability of causal relationships, the types of relationship that are at stake when we study the effects of a new drug to delay the onset of Alzheimer's disease, the effects of a "tough love" program for teaching employment skills to the long-term unemployed, or the effects of physical exercise on cognitive skills in the elderly. Following the work of Campbell and Stanley (1966), Cronbach (1982), and Cook (1993; Cook & Campbell, 1979), I will distinguish three types of empirical (i.e., data-driven) generalizations. The first concerns inferences to target populations, classes, or universes (e.g., the population of elderly, the class of retirement homes, the universe of cognitive skills). The second involves generalization across populations or sub-populations (e.g., public and private retirement homes, men and women, verbal and numeric skills). The third involves extrapolation and interpolations about novel universes. After presenting some of the traditional approaches to justifying generalized inferences, I will review five principles proposed by Cook (1990, 1993) to strengthen empirical generalizations, illustrate their application in the context of meta-analysis, and discuss the conditions under which valid generalizations are most likely to emerge. The chapter argues that strong generalizations are rarely – if ever – possible based on single studies. Instead, generalizations are best justified by programmatic reviews of findings from many studies, the type of reviews to which carefully conducted meta-analyses can make significant contributions.

8.2 THE DIFFERENT MEANINGS OF GENERALIZATION

The term "generalization" has many meanings and connotations in everyday life and in scientific discourse. Some of these meanings are briefly reviewed in the following.

8.2.1 Crisp and Fuzzy

In everyday discourse, "generalization" refers to a proposition that applies to a large number of instances of a class or group (Webster, 1986). For instances, "It never rains in Southern California." could be interpreted in a crisp manner to mean that there are zero days with precipitation south of Santa Barbara, CA. In addition, "generalization" has the connotation of "vague" or "fuzzy" in that the proposition may not always apply in exactly the same way to each and every member of a group or class (Zadeh & Yager, 1987). That is, that fact that San Diego is in Southern California, receives on average 10 inches of rain per year, and received about 1/2 inch of rain last night does not necessarily disprove the "fuzzy" generalization that one should not expect rain on prototypical days in southern California.

8.2.2 Inductive and Deductive

Generalizations are often the result of inductive inferences, in which general statements are made based on specific observations. Watching a few episodes of Baywatch may lead many TV viewers to the conclusion that, in general, residents of Southern California are very attractive, athletic, and adventurous.

Generalizations may also be the result of deductive inferences in which a general proposition leads to more specific conclusions. Given the general wisdom that Westphalians are stubborn but produce excellent ham, one would expect to find headstrong persons and excellent hams throughout Westphalia's large industrial area around Essen, university towns like Münster, its hamlets along the Dutch border, and its expatriates in Wisconsin.

8.2.3 Logical, Empirical, and Theoretical

In formal logic the validity of an inference depends entirely on its form or structure and not on the subject matter (Groeben & Westmeyer, 1975). A valid inference is one in which a proposition (e.g., a general conclusion) follows with strict necessity from a set of premises (e.g., syllogism). The deduction of the conclusion from the premises must follow the formal rules of logic. For instance, given the premises that "All psychotherapies are effective" and that "Interpersonal therapy is a form of psychotherapy", it follows that "Interpersonal therapy is effective". As soon as the truth of the premises has been established, the validity of the argument is ensured based on the structure of the argument alone. Formal logic is concerned with inference forms rather than with the particular instances.

In contrast, empirical generalizations – the topic of this chapter – are informed primarily by patterns of observations made in particular instances, with inference forms playing a secondary role (Groeben & Westmeyer, 1975). The logic underlying empirical generalizations is closely related to Popper's falsificationist (Popper, 1959, 1972) approach as applied in quasi-experimental designs where plausible alternative hypotheses are identified and ruled out (Cook & Campbell, 1979). The goal of empirical generalizations is not to establish truth but to explore the dependability of generalization claims by subjecting them to falsification tests. That is, empirical generalizations are always tentative and approximate. The more a generalization has been subjected to credible empirical falsification tests, the stronger the belief in its validity.

Empirical generalizations also have to be distinguished from the class of general propositions that make up a theory. Theoretical generalizations are law-like statements that do not directly apply to the empirical world. Instead they rely on theoretical constructs, abstractions and simplifications of complex empirical phenomena, and idealized conditions. Theoretical generalizations present an ideal model of the real world; empirical generalizations present an empirical model of the real world.

8.2.4 Universal and Specific

Universals are empirical generalizations that are of near complete generality (Abelson, 1995) such that they apply to all humans or all humans of a certain type (e.g., retirees taking physical exercise classes). While there are more universals in the natural sciences, there are many near universals in the social and behavioral sciences as well. Examples of nearly universal findings in psychology include: (a) Limitations of short-term memory cause humans to chunk information into groups of no more than about seven items (i.e., Magic Number 7 ± 2); (b) Language acquisition only occurs in humans if they have had exposure to a language community during a critical period in infancy; (c) When deciding among courses of action with equal objective payoffs humans are risk-averse and select the least risky option; and (d) The emotional interpretation of prototypical facial expressions. While exceptions exist to all these near universals, these exceptions are rare, highlighting that they are "exceptions to the rule", thereby corroborating the existence of the rule.

Generalizations do not have to be universals. Instead, propositions can be phrased at different levels of generality, ranging from near universals to statements identifying specific circumstances under which the proposition holds. The higher the level of generality, the broader the range of circumstances across which the proposition presumably holds.

8.2.5 Transfer, Extrapolation, and Analogs

Transfer refers to a form of generalization where observations made in one condition are extrapolated to another. In learning theory (Estes, 1978; Hommel & Prinz, 1997), transfer is said to have occurred when a subject who learned to respond to a particular stimulus (e.g., a 440 Hz sound) responds as well to similar stimuli beyond the original conditions of training (e.g., 597 Hz, 293 Hz). As differences between two conditions increase, the effects of generalization decrease until there may be no transfer from one situation to another. Alternatively, the more the two situations have in common, the greater is the amount of predictable transfer.

The transfer view of generalizability underlies training approaches in areas where learning on the job can be prohibitively expensive, outright dangerous, or inappropriate for practical and ethical reasons (Cormier & Hagman, 1987). For instance, training pediatric surgeons on critically ill infants to perform a new form of heart catherization, training navy F-14 fighter pilots in actual war situations for combat missions, or training operator personnel of nuclear power plants in actual catastrophic accidents are ethically indefensible. In all of these examples, an important training component takes place on analog counterparts of the actual situations, such as animal models, flight simulators, or analog control rooms. These analogs are designed to maximize the amount of positive transfer (i.e., training that facilitates actual performance), and minimize negative transfer (i.e., training that hinders actual performance).

The transfer view of generalizability is also at the core of *technology transfer* models that aim at facilitating the transition of research findings obtained under laboratory conditions to commercial applications (National Academy of Sciences, 1997). A good example of the implementation of the technology transfer model is the approval process that guides the development of new pharmaceutical products in the U.S.A. (U.S. Food and Drug Administration, 1998). Experimental new drugs are first tested in preclinical studies for safety and efficacy, involving cell cultures, computer models, or animals. If the Food and Drug Administration (FDA) review panels come to the conclusion that findings from the lab are likely to extrapolate to humans, approval is granted for Phase I clinical studies in humans. These are short-term studies on small samples, focusing on safety, and often involve healthy subjects. If these initial studies demonstrate a drug's safety, Phase II studies follow, in which short-term efficacy and drug safety are investigated in larger samples. If Phase II studies continue to demonstrate a drug's safety and efficacy, large-scale Phase

III clinical trials are conducted with a focus on drug dosage, long-term effectiveness, and drug safety. Data collected in the preclinical and clinical trials are again reviewed by FDA expert panels to approve, to request additional studies, or to deny approval of a new drug. Following the approval of a new drug, monitoring systems are put in place to detect adverse reactions and to investigate quality control. The FDA estimates that only 5 of 5,000 compounds entering preclinical testing make it to human testing. Of those, only 1 in 5 are eventually found to be safe and effective and approved for marketing.

8.2.6 Replicability and Robustness

The better research findings replicate across different conditions, the more general the effect is said to be. Robust empirical findings suggest broad main effects of interventions, the type of effects that are particularly useful for policy decisions affecting large and diverse constituencies (Abelson, 1995).

If research findings are replicable within conditions but vary across conditions, interaction effects are present that moderate the direction or magnitude of an effect. Such interactions help identify the boundaries of generalizability. Of particular interest are conditions that reverse the direction of an effect (i.e., qualitative interaction) as is the case when physical exercise lowers blood pressure in some groups but increases blood pressure in others.

8.2.7 Fixed and Random

While robust main effects promise broad generalizability, Abelson (1995) points out that there is a catch. If main effects were investigated based on a limited number of fixed levels (e.g., 7 hours vs. 0 hours of vigorous exercise per week), a disclaimer is necessary stating that the generality is limited to the specific levels represented by the factor. To avoid the limitations of a fixed factor, contexts across which one intends to generalize should be considered as random factors with many different levels from which a sample is being investigated to draw inferences about the whole (e.g., many different durations, types, and intensities of physical exercise).

While a random effects model of contexts will be desirable in many generalizability situations, there are exceptions. Sometimes, researchers deliberately constrain their inferences to particular fixed levels on some factor and random levels on some other factors. For instance, a new psychotherapeutic intervention may rely on a highly standardized treatment manual, administered in controlled inpatient hospital settings for a specific eating disorder (e.g., binge eating). Similarly, the effects of a new drug may be of interest at limited and fixed dosages and for a carefully selected subset of persons suffering from a specific illness.

8.3 A FRAMEWORK FOR EMPIRICAL GENERALIZATIONS

The following framework relies on the work of Brunswik (1947), Campbell and Stanley (1966), Cronbach (1982), and Cook (1990, 1993) on the validity of causal inferences in field settings.

8.3.1 Representative Designs

Brunswik (1947) and Hammond (1948, 1951) were among the first psychologists to raise objections against studying macro-level behaviors with experimental methods under laboratory constraints. Brunswik (1952) argued that when studying behavior at a macro-level "... care must be exercised not to interfere with naturally established mediation patterns." (p. 26). Such an approach calls for research designs that are representative of the natural conditions in which the behavior takes place, that is, designs which Brunswik referred to as having situational representativeness, "naturalness, normalcy, 'closeness to life' " (Brunswik, 1952, p. 29). Clearly, at issue are designs with ecological or situational validity.

For Hammond (1948, 1951) and Brunswik (1952), representative designs lead to generalized statements if a reference class or universe has been specified from which situations are sampled and about which inferences are sought. To achieve a representative design requires not only representative sampling of individuals but also sampling the situational circumstances under which a person functions outside of the research laboratory. This includes stimuli or interventions, responses or outcomes, and settings.

8.3.2 Domains About Which Generalizations May Be Desired

Campbell and Stanley (1966), Cronbach (1982), and Cook and Campbell (1979) have identified five entities about which generalizations may be desired. First are persons or, more generally, the units (U) to which treatments have been assigned. Units may consist of individual humans, animals, or cells as well as larger aggregates such as families, schools, neighborhoods, or states. A second entity is treatments (T), for which a specific operationalization was implemented in a study (i.e., cause constructs). A third entity is outcomes (O) of which specific operationalizations were used to measure effects of interest (i.e., effect constructs). A fourth entity is settings (S), referring to the social and physical environment in which the study takes place. A fifth entity is time (Tm), indicating the historical context in which the study takes place[1].

Each of these domains can involve universes at different levels of generality. For instance, the domain of U may consist of U.S. residents 60 years and older

[1] Note that Cronbach subsumes time in his definition of setting, a distinction with minor influence for the discussion that follows. I will keep with Cook and Campbell's (1979) distinction to better reflect distinct generalizability questions with regard to the social/physical and historical contexts.

or the subset (i.e., sub-U) of residents living in California with annual household income between $20,000 and $30,000 who are registered as independent voters. Similarly, the domain of T may consist of all types of vigorous physical exercises or the subset (i.e., sub-T) of exercises involving stationary bicycles.

A domain may consist of a few fixed levels or an infinite number of random levels. For instance, in some clinical trials of new drugs great efforts are made to control (i.e., fix) as many components of a study as possible. Research protocols are designed and their implementation is carefully monitored in different sites, prescribing in detail the specific characteristics of subjects to be recruited, specific levels of drug dosages to be administered, specific end points to be measured, and specific settings in which treatments are administered and patients are monitored. The goal of these studies is to collect evidence about a specific type of patients, specific dosage levels, specific outcomes, and in closely controlled settings.

In other studies, the aim is to draw inferences about universes consisting of large number of instances. In these situations, the researcher samples a subset of instances to represent the entire domain. For instance, a new algebra curriculum may be tested in public and private schools, with junior and senior instructors, in rural, suburban, and urban areas to draw inferences about the curriculum's effectiveness across a wide variety of school and students.

8.3.3 Generalizability Questions

In their classic work on quasi-experimentation and the validity of causal inferences in field settings, Cook and Campbell (1979) distinguish two major generalizability questions. The first question concerns *generalizations to* well-explicated target domains. This question is invoked when we ask whether a particular sample of retired persons allows valid inferences about the target population consisting of all retired persons. In Cronbach's notation, this question asks whether we can draw inferences from utos to UTOS where u, t, o, s designate the samples of U, T, O, S realized in a particular study. These concepts will be elaborated on later. The first question is most closely associated with inductive probabilistic inferences from samples to populations.

The second question concerns *generalizations across* well-explicated subdomains (i.e., inferences about sub-UTOS). This question is invoked when we ask which different populations or subpopulations (e.g., rural vs. urban vs. suburban; public vs. private; 8th grade vs. 9th grade vs. 10th grade) have been affected by an intervention. The second question is most closely associated with deductive inferences, in which the robustness of a general proposition is investigated across different context conditions. In Cronbach's notation, this question asks whether we can draw inferences from utos to different subsets of UTOS.

Cronbach argues that there is a third generalizability question, which is of particular concern in applied areas of research. This question involves *generalizing from the specific samples and the universes they represent to novel universes* not yet studied. For instances, having found that co-payments for doctor's visits reduce the number of unnecessary visits in San Diego, CA, will co-payments have the same effect in Münster, Westphalia? In Cronbach's notation, this question concerns inferences from utos to *utos, inferences from the domain of observation to the domain of application. The third generalizability question is closely related to the transfer view of generalizability as it clearly invokes an extrapolation from conditions in which a research finding was investigated to related yet new and distinct conditions.

8.3.4 Justifying Empirical Generalizations

8.3.4.1 Complete Causal Explanation In the experimental tradition of the social and behavioral sciences, generalizations are justified through complete explanation, that is the complete understanding of the causal conditions and mechanisms that bring out a phenomenon. The assumption behind this belief is that when we understand how or why a phenomenon occurs, we can recreate that phenomenon wherever and however its causal ingredients can be brought together (Bhaskar, 1978). This is why causal explanation is often considered the "Holy Grail" of science and the scientific method the path leading to it.

Take for instance the recent approval of thalidomide (alias Contergan) for the treatment of uncontrolled blood vessel growth and severe immuno-modulated diseases. In the 1950s, Chemie Grünenthal, a German pharmaceutical company, developed a sedative called thalidomide so harmless to rodents that an LD50 could not be established (i.e., lethal dose 50 is a measure of acute single exposure toxicity; it indicates the dosage at which 50% of the animals die). The causal mechanisms set in motion by thalidomide were not completely understood, a situation not uncommon even today in many popular drugs (e.g., aspirin, ibuprofen). In the late 1950s and early 1960s, at least 10,000 pregnant women in 46 countries took the sedative in their first trimester, eventually giving birth to infants with missing or stunted limbs. In the early 1960s McBride (1961), Lenz (1962), and Pfeiffer and Kosenow (1962) reported the association between maternal thalidomide usage and limb defects in babies, leading to a world-wide ban. It was not until 30 years later, that an explanation was found for how this potent human teratogen caused missing and stunted limbs.

D'Amato, Loughnan, Flynn, and Folkman (1994) discovered that thalidomide inhibits angiogenesis (i.e., blood vessel growth) in rabbit corneas, changes similar to those found in the deformed limb bud of thalidomide-exposed embryos. It is the ability to inhibit angiogenesis that most likely caused limb defects in babies after maternal thalidomide usage. The causal understanding of how thalidomide works is now being applied to treat conditions characterized by uncontrolled angiogenesis, including diabetic retinopathy and macular degeneration in populations not at risk of becoming pregnant (e.g., Verheul, Panigrahy, Yuan, & D'Amato, 1999).

8.3.4.2 Sampling Theory In the observational and correlational traditions of the social and behavioral sciences, causal generalizations are often justified through the correspondence between samples (or cases, instances, exemplars, etc.) and the populations (or universes, constructs, categories, classes, etc.) they are meant to represent. The assumption behind this belief is that causal relationships must be easiest to reproduce under the same or similar circumstances they were originally demonstrated. Carefully selecting the specific conditions under which causal effects are demonstrated (e.g., subject characteristics, outcome measures) may allow to approximate important larger classes in which the causal effects hold.

To justify inferences from samples to populations, statistical sampling theory has long played a crucial role in survey research and quality control in industry (Kish, 1965). The crucial element in sampling theory involves selecting units (e.g., persons, hospitals, observers, therapists) with known probability from some clearly designated universe so as to match the sample and population distributions on all (measured and unmeasured) attributes within known limits of sampling error. That is, if one can demonstrate that co-payments for doctor's visits causally reduce unnecessary visits in a random sample of doctors' offices belonging to a particular HMO, the effect in the entire population of HMO subscribers can be estimated.

Note that valid causal explanations are neither a necessary nor a sufficient condition for causal generalizations based on sampling theory. Instead, the causal generalizations based on sampling theory may be a particularly useful tool when complete causal explanations are not available. Similarly, causal explanations may be achieved based on careful experimentation on a few specimens and under highly controlled conditions with little consideration given to sampling theory. For instance, probability sampling did not play a significant role in the original development of thalidomide as a sedative, in the discovery of its devastating side effects, or in the recent approval for new medical indications. However, a case could be made that the thalidomide tragedy could have been reduced had more careful attention been given to sampling principles, including the definition of the universes of persons and outcomes about which inferences are desired and the selection of exemplars from these universes.

8.3.4.3 ***Campbell's Models for Increasing External Validity*** Campbell and Stanley (1966) distinguished internal validity from external validity to highlight two distinct inferences about the validity of experiments in field settings. Internal validity refers to the approximate validity with which we infer that the relationship between the manipulated cause and the measured effect is causal. External validity refers to inferences about the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of cause and effect, and across different persons, settings, and times.

Campbell and Stanley (1966) and later Cook and Campbell (1979) acknowledge that internal validity by itself is concerned only with the specific circumstances of how a presumed cause was manipulated and how a presumed effect was measured. Clearly, external validity is necessary before we can expect to replicate a causal relationship in a different sample of persons, with different manipulations of the presumed cause, different operationalizations of the outcome, or in different settings.

To strengthen external validity in an individual study requires implementing strategies to better represent classes of persons, treatments, outcomes, or settings. If feasible, such strategies include random sampling (e.g., drawing a sample of affectively valenced words from a list of all such words in the English language), impressionistic samples of modal instances (e.g., selecting prototypical public and private schools from rural, urban, and suburban areas), or the deliberate sampling for heterogeneity (e.g., recruit diverse participants with respect to gender, age, income, ethnicity).

While some of these strategies may work in some studies and for some of the entities about which generalizations are desired, they are unlikely to work in most studies and for all generalizations of interest. With few exceptions, individual studies are often constrained by the unique selection of persons, settings, and times, rendering it impossible to draw generalized inferences about larger classes. In many studies, researchers do not have adequate access, budgets, or time to carefully select probability samples from target universes – if meaningful sampling frames for such target populations exist at all. Instead, researchers often select fixed levels from these populations, limiting inferences to such fixed instances. Or, researchers rely on convenience samples, in which cases inferences about a target populations can not be made based on sampling theory. Clearly, external validity is the Achilles' Heal of causal inferences based on an individual study.

8.3.4.4 Cronbach's Model-Based Reasoning for Justifying Internal and External Inferences Cronbach (1982; Cronbach, Nageswari, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) made two significant contributions to our understanding of generalizability. The first concerns the dependability of observations known in the literature on measurement theory as generalizability theory or G-Theory. G-Theory provides a framework for designing and investigating reliable observations by reinterpreting classical reliability theory (Nunnally & Bernstein, 1994) as a theory regarding the adequacy with which one can generalize from a sample of observations to a universe of admissible observations. The universe of admissible observations consists of observations that are interchangeable for the purposes of making a measurement decision. Observations are "dependable" or "generalizable" if they permit accurate inferences about the universe of admissible observations.

Cronbach's second contribution concerns the generalizability of program evaluations. Similar to G-theory, Cronbach defines a domain of admissible operations about which an investigator asks questions and would like to draw inferences. This domain consists of subjects or units (U), interventions or treatments (T), procedures for collecting data on outcomes (O), and the historical and cultural conditions or settings (S). To draw inferences about UTOS, an investigator collects data on instances of the various domains, referred to with lower case letters u, t, and o. Because researchers have little control over the social and historical context of their research, they can rarely sample instances from S.

According to Cronbach's model, *internal inferences* are involved when making statements about UTOS on the basis of observations on utoS. Questions about the trustworthiness of these inferences are questions about internal validity. Note that Cronbach's internal validity is identical to Campbell's external validity when generalizations to a particular universe are of interest.

In addition to statements about UTOS, investigators and as consumers of research may be interested in domains that are different from the original. Cronbach calls this the domain of application and refers to it as *UTOS. The second generalizability question thus concerns inferences from utos to *UTOS. Statements about *UTOS involve external inferences or extrapolations if we would like to draw inferences about subjects populations, treatments, or outcomes not included in the original study. According to Cronbach's model, questions about the trustworthiness of external inferences are questions about external validity. Note that Cronbach's external validity concerns a generalizability question that Campbell did not consider in his model of external validity.

To justify internal and external generalizations, that is inferences from utos to UTOS or and from utos to *UTOS, Cronbach proposes reasoning by means of models. To justify internal inferences, models are constructed that simulate specific research problem. Models may be descriptive, explanatory, physical, mathematical, or logical, including the blueprints of an architect, the scale model of an engineer, the micromediational model of a microbiologist, for the mathematical model of a survey researcher. Conclusions are drawn in the model and then translated to the real world. Whether conclusions about UTOS are trustworthy depends on the extent to which the model is complete and credible.

Cronbach (1982) describes inferences about *UTOS "as a multi-track, if not trackless process" (Cronbach, 1982, p. 166) because different types of evidence and reasoning have to be combined. Any conclusion about *UTOS must be informed by the differences and similarities between *UTOS and UTOS. Clearly, the more *UTOS and UTOS differ the more has to be filled in to bridge the gap through complementary evidence and credible models to permit trustworthy projections. In general, external inferences about *UTOS are associated with considerably more uncertainty than statements about UTOS.

Cronbach's model-based justifications of generalized inferences include and go beyond the traditional justifications provided by sampling theory or causal explanation. Models may include the complex sets of mathematical equations used by economists to project the effect of increasing oil prices on inflation and unemployment rates. They may also involve informal heuristic models in which many of the premises may not be explicit and in which judgment and formal reasoning have to be combined. Regardless of the model, the credibility of the generalized inference rests on the extent to which the relevant research community accepts the assumptions it is build on. As Cronbach points out, the acceptance of generalized conclusions rests as much on social psychological processes as it rests on the sheer strength of the empirical evidence with which different parts of a model can be supported.

8.3.4.5 Cook's Five Principles for Strengthening Causal Generalizations

Building on Campbell and Cronbach's work, Cook (1990, 1993) set out to examine how researchers have achieved generalizable causal relationships in the absence of strong causal models and probability sampling. While Cronbach provides a theoretical account of how generalizable claims are substantiated, Cook proposes five principles that researchers use to strengthen claims about the generalizability of causal claims. Cook's work is particularly interesting because it points to strategies and conditions that can be applied in planning individual research studies and be helpful when synthesizing findings from many individual studies.

The Principle of Proximal Similarity. Campbell (1969) introduced the notion of proximal similarity in the context of construct validity. In the context of generalizability, Cook (1990) expands its definition to refer to the correspondence in manifest descriptive attributes between a class of persons, settings, causes, outcomes, and times about which generalizations are sought and the instances based on which empirical evidence about a causal relationship are available. The similarity is proximal because samples and universes match in observable characteristics and not necessarily in any of the more latent explanatory components that link a cause to an effect (Cook, 1990). Proximal similarity is clearly in the spirit of Brunswik's (1956) "situational representativeness".

Demonstrating proximal similarity to the critical standards of the research community is the first necessary condition for generalizing to target universes (i.e., first generalizability question). Proximal similarity is achieved by explicating and then matching the multidimensional content of the classes and instances involved in the generalization. But matching cannot be achieved on all components. Therefore, matching is most importantly achieved with those components that theoretical analysis suggests are central to the construct description.

Cronbach's notion of a domain of "admissible observations" may also be used to argue for proximal similarity. The idea is that a convenience sample of persons, treatments, and so forth may be considered representative if the instances included in the sample and the instances not included in the sample are equally acceptable or exchangeable. Shavelson and Webb (1981) even argue that under these conditions a sample should be considered random. What defines a domain of admissible or exchangeable instances depends on the set characteristics a researcher considers substantive irrelevant (i.e., exchangeable) and the set of characteristics deemed substantively relevant (i.e., prototypical). The latter defines the necessary conditions and the former the unnecessary (i.e., irrelevant) conditions of group membership.

The Principle of Heterogeneous Irrelevancies. A causal relationship will be easier to generalize if it has been replicated in multiple studies especially if these replications involved different research teams, multiple populations, in multiple settings, with multiple implementations of treatments, and multiple outcome measures. Of interest are replications that are proximally similar with respect to conceptually relevant components but differ in all conceptually irrelevant ones. The principle of heterogeneous irrelevancies can strengthen causal generalization by examining whether the cause-effect relationship under investigation is robust or contingent upon a particular irrelevancy or set of irrelevancies. This is exactly what Cronbach et al. (1972) argue when generalizing from a sample of observations to a universe of admissible observations. In synthesizing findings across irrelevancies, we ask whether the irrelevancies make a difference and whether the causal relationship is obtained *despite* the irrelevancies.

The principle of heterogeneous irrelevancies provides a second necessary condition for generalizing to target universes (i.e., first generalizability question). Findings regarding the benefits of physical exercise in the elderly become trustworthier if they are robust across different type of exercise, different populations of elderly, for different levels of functioning, in different settings. As part of the new drug approval process, the FDA requires preclinical trials to involve at least two animal species to make heterogeneous the presumably irrelevant aspects of the genetic make-up (U.S. Food and Drug Administration, 1998). Clinical trials of new drugs have to be studied in different age and gender groups to determine the robustness (or lack thereof) across these groups. Perhaps the most elaborate application of this principle can be found in meta-analyses of psychotherapeutic interventions, demonstrating the robustness of effects (in causal direction) across a wide variety of different irrelevant characteristics of the researchers, the research design, the intervention, patients, and so forth.

The Principle of Discriminant Validity. The principle of discriminant validity calls for investigations that disentangle the many constituent components of a setting, cause, population, outcome, and time period, to determine the extent to which these components are necessary, sufficient, or irrelevant to the causal relationship under investigation. Through experimental manipulation and observational studies, the goal is to investigate treatment effects in subpopulations, in different settings, with different treatment components, and across different outcome constructs to identify the causal efficacious conditions and discriminate them from related though inefficacious conditions.

This approach does not help, however, if the variations in subpopulations, settings, and so forth are limited such that they all share a common bias. For instance, when all subjects are male or all outcome measures rely on self-report, treatment effects are confounded with gender and assessment method. To apply the principle of discriminant validity, treatment effects have to be studied across populations, treatments, outcomes, and settings with many levels, representing the range across which generalizations are desired.

Moderator effects play an important role in characterizing the boundaries of generalizability and identifying the conditions under which the direction or strength of a relationship may vary. If this moderator involved a hypothesized substantive irrelevancies, its status now changes from that of a substantive relevancies and attempts should be made to better understand the role of this theory-relevant construct. Investigations of moderator effects are closely associated with the second generalizability question when generalizations across different UTOS to sub-UTOS are of interest.

Discriminant validity is a necessary condition for generalizing across universes. In combination with proximal similarity and heterogeneous irrelevancies, discriminant validity strengthens generalizations by identifying the limits of generalizability and the conditions under which effect changes in sign or magnitude. The principle of discriminant validity is invoked when researchers study dose-response relationships, examine treatment effects across different populations, or distinguish target outcome from side effects.

The Principle of Causal Explanation. While causal relationships are concerned with establishing whether a causal link exists, causal explanations are concerned with identifying *how* or *why* a causal connection occurs. They involve specifying the full set of conditions promoting the cause-effect connection, which often entails identifying the mediational forces set in motion when the treatment varies and without which the effect would not occur.

Causal explanations strengthen empirical generalizations. However, they are not sufficient nor are they necessary conditions for generalizations. Cook (1990, 1993) concludes that the role of causal explanations for justifying generalizations may be overrated. He argues that given the paucity of strong causal explanations and the nature of many problems investigated in the behavioral sciences, it is often unrealistic – if not unethical – to expect and wait for complete causal explanations before attempts are made at causal generalizations. For instance, understanding the micro-mediating processes of a new drug on a molecular level has great significance for making predictions about potential effects in humans. However, it is neither a necessary nor a sufficient condition for the FDA to consider a drug safe and effective. The FDA approval rests primarily on the body of empirical evidence regarding the drug's safety and effectiveness for a particular indication in particular populations and at particular dosage levels. At the same time, a complete causal explanation for the operation of a drug is not sufficient to justify that it is safe and effective for marketing. The recent FDA approval of thalidomide was only given after comprehensive clinical trials despite the fact that the causal mediating mechanisms are quite well understood.

The Principle of Empirical Interpolation and Extrapolation. Populations, treatments, outcomes, settings, and times can vary along many different dimensions. Some of these dimensions are quantitative, in which case special opportunities arise for justifying certain generalizations. Examples for such quantitative dimensions are age, income, weight, family size, treatment dosage or duration, and quantitative outcomes. If we consider these dimensions as fixed factors and collect data at strategically spaced levels (Abelson, 1995), we create the opportunity to describe the quantitative relation between effect and dose, duration, age, income, and so forth. Assuming valid characterizations of these quantitative relationships, we can derive interpolations and extrapolations about levels of these factors that have not yet been studied. This form of empirical extrapolation and interpolation may strengthen inferences about novel conditions for which no empirical data are available. Thus, empirical interpolations and extrapolations are at the center of the third generalizability question (i.e., generalizations to novel universes).

Interpolation is involved when characteristics can be ordered along a quantitative dimension (e.g., dosage) and when inferences about this characteristic are desired at a level that falls between two known levels. For instance, this is the case when we infer the effects of a drug dosage at 450 mg based on individual studies of the effects at 200, 300, 400, 500, and 600 mg. The narrower the gap and the more data points are available below and above the gap to be interpolated, the more confident are we about our interpolation because the dose-effect relationship is less likely to change abruptly over the interpolated gap.

A similar rationale holds for extrapolations, where we have studied treatment effects at levels 5, 6, 7, 8, 9, 10 and are interested in generalizing to treatment effects at levels 2 and 4, or 12 and 20. Again, the shorter the gap across which we extrapolate and the wider the range of levels across which we studied the relationship, the more confident we are in the extrapolation inferences. Moreover, the wider the range of levels across which we have studied the relationship, the more confident we are that we have identified the proper mathematical model to make the extrapolation (e.g., linear or logarithmic). The shorter the extrapolation leap, the less likely it is that the relationship between level and effect does not change abruptly.

The extrapolation inference will always be weaker than the interpolation inference because we have collected evidence regarding the nature of the relationship from only one direction. From this perspective, interpolations can be sought of as two extrapolations that can be pooled to yield a better estimate.

Interpolations and extrapolations are model-based predictions, whose validity hinges on the assumption that the model holds in between the levels to be interpolated and at the levels to be extrapolated. There are many examples in the natural and behavioral sciences where such assumptions are patently false and relationships change abruptly or take on new qualitative forms. For instance, the physical properties of water change dramatically at specific temperatures. Many pharmaceutical compounds have beneficial effects across a certain range of dosage but may have no effect below and lethal effects above that range. Similarly, in certain problem solving tasks motivation and performance are positively related up to a point at which increases in motivation lead to a decline in performance.

8.4 COOK'S PRINCIPLES APPLIED TO META-ANALYSIS

As a tool for "communal testing of generality" (Abelson, 1995), meta-analysis holds great promise for justifying generalized inferences regarding all three generalizability questions. Matt and Cook (1994) have argued that the generalizability of meta-analytic inferences is particularly justified when

a) the universes about which generalizations are desired are well matched by the instances represented in individual studies (i.e., proximal similarity),
b) individual studies share substantively relevant features but are heterogeneous with respect to irrelevant features (i.e., heterogeneous irrelevancies), and
c) studies can be disaggregated to investigate substantively meaningful subclasses (i.e., discriminant validity).

8.4.1 Meta-Analysis and the Principle of Proximal Similarity

Psychotherapy outcome studies are perhaps the best reviewed body of empirical research using meta-analytic methods (Matt & Navarro, 1997). Following Smith and Glass' (1977) initial meta-analysis of about 500 psychotherapy outcome studies (see also Smith, Glass, & Miller, 1980), more than 50 additional meta-analyses had been conducted by 1992, with many more since then. Glass and colleagues set out to investigate whether psychotherapy in general is beneficial. In the framework presented above, this question implies the desire to generalize to the universe of interventions labeled psychotherapy (T), the universe of persons receiving treatment (U), the universe of settings in which the treatment takes place (S), the universe of outcomes used to assess effects (O), and the historical period during which psychotherapy has been practiced (Tm). Smith and Glass (1977) estimated that psychotherapy treatment group patients did about eight-tenths of a standard deviation better on the outcome variables than did patients in the control groups. Overall, the empirical generalization appears warranted that psychotherapy works.

How can such a general conclusion be justified? The justification begins with investigating the proximal similarity between target universes and instances included in the meta-analysis. Evidence has to be generated that the broad universes of UTOSTm were well represented by samples included in the meta-analysis. The goal is not an exact match or probabilistic representation as sampling theory suggests. Instead, the goal is to achieve an approximate match on prototypical components with multiple operational representations. Or, in Cronbach's (1982) terms, one has to argue that the instances of UTOSTm included in the meta-analyses are exchangeable with the instances not included. In Smith and Glass' meta-analysis (1977), "psychotherapy" included a variety of orientations and techniques such as psychodynamic, behavioral, cognitive, interpersonal, hypnosis, bibliotherapy, eclectic, and others. Similarly, outcomes included a multitude of measures, ranging from global indices of adjustment to frequency counts of a specific symptom and from standard trait inventories to ad hoc therapist ratings. It appeared that key prototypical characteristics of "psychotherapy" were represented in the sample of studies included in Smith and Glass' meta-analysis.

8.4.2 Meta-Analysis and the Principle of Heterogeneous Irrelevancies

Once a case has been made for proximal similarity, meta-analysts have to generate evidence that a causal connection is not completely confounded with any specific characteristic of u, t, o, s, tm. This calls for the application of the principle of heterogeneous irrelevancies. The greater the number of irrelevancies across which primary studies differ, the greater the chance that a causal connection is not completely confounded. The assumption that substantive irrelevancies are heterogeneous should not be made lightly. It has to be based on evidence that mono-operation biases are not present across the domains of which generalizations are desired.

Lack of heterogeneity was not a problem in the Smith and Glass meta-analyses. Smith and Glass coded primary studies to collect data on many substantively relevant and irrelevant study characteristics, including year and source of publication, professional affiliation of authors, age, gender, socioeconomic status of participants, type and reactivity of outcome measures, type of control conditions, sample size, duration of therapy, experience of therapist, recruitment of subjects, and setting in which therapy took place.

Overall, the heterogeneity in the sample of studies appeared to match the heterogeneity of the domain about which inferences are desired. That is, psychotherapy outcome studies are conducted by many different research teams, researchers with training in different disciplines and with different professional affiliations. Researchers use different approaches for recruiting subjects, implementing treatments, and measuring outcomes. Similarly, psychotherapy is conducted across a wide range of settings, including private practices, schools, community mental health centers, and university-affiliated hospitals.

Within specific subclasses of treatments, heterogeneity was reduced, giving rise to potential mono-operation biases. For instance, some of the subclasses of interventions relied more heavily on small sample sizes, student volunteers, school settings, short therapies, and certain types of outcome measures. Heterogeneity in studies is welcomed if it matches the heterogeneity of the domain about which inferences are desired. However, any restriction would limit conclusions regarding the generalizability, as was the case with certain subclasses of psychotherapeutic interventions (Matt & Navarro, 1997).

8.4.3 Meta-Analysis and the Principle of Discriminant Validity

While proximal similarity and heterogeneous irrelevancies are at the center of generalizations to target domains (i.e., first generalizability question), the principle of discriminant validity plays a key role when generalizing across subdomains. Rather than lumping heterogeneous studies together, meta-analysts can test whether the heterogeneity in treatment effects is larger than expected due to chance alone (Hedges & Olkin, 1985). Equally important is the decision whether to rely on a fixed, random, or conditional random effects model (Hedges, 1994; Hedges & Vevea, 1998). This decision is influenced by whether a researcher is interested in drawing inferences about a few clearly defined (i.e., fixed) subclasses of a domain or to the entire domain consisting of a large number of admissible parts.

Depending on the statistical model chosen and sample size permitting, heterogeneous domains of U, T, O, S, Tm may be disaggregated to identify more homogeneous subdomains. This starts the exploration of potential interaction effects, that is, substantively relevant and substantively irrelevant characteristics that moderate the size or direction of treatment effects. For instance, such analyses may indicate that different types of interventions or different outcome constructs (e.g., a substantively relevant heterogeneity) are associated with different effect sizes but that treatment effects are robust with respect to the type of setting or subject recruitment.

The principle of discriminant validity is applied in meta-analyses when studies are stratified to investigate moderator conditions (e.g., gender, treatment types) or to distinguish important cognate constructs from each other (e.g., functional disability, life satisfaction, self-esteem, symptoms, adjustment). As mentioned above, Smith and Glass' meta-analysis spawned a large number of additional meta-analyses on psychotherapy effects (Matt & Navarro, 1997). The purpose of these additional meta-analyses was to explore whether psychotherapy effects generalize across different types of interventions, outcomes, settings, and populations. While there is evidence that certain conditions moderate the magnitude of psychotherapy effects, none of the meta-analyses identified conditions associated with harmful effects (Matt & Navarro, 1997). That is, the beneficial effects of psychotherapy generalize across wide range disorders, types of interventions, outcomes, settings, and time periods.

The principle of discriminant validity was also applied in meta-analytic investigations of the placebo effect in psychotherapy (Matt & Navarro, 1997). At issue is the question whether a treatment group, which is presumed to be receiving psychotherapy, is actually receiving both psychotherapy and a host of nonspecific placebo interventions. The latter include, for example, the mere attention given by a caring person, the expectation of improvement brought about simply by being seen by a mental health professional with credentials for dealing with psychological problems, or by the simple fact of talking with another human being about the problem. Such placebo effects have proven to be so powerful in medicine that they need to be controlled by using double-blind designs and introducing placebo control groups.

To examine the role of placebo effects in psychotherapy, at least eight meta-analyses compared experimental studies in which patients in psychotherapy were compared against patients who received placebo treatment that did not include the presumed active therapy ingredient (Bowers & Clum, 1988; Casey & Berman, 1985; Clum, Clum, & Surls, 1993; Kazdin, Bass, Ayers, & Rodgers, 1990; Landman & Dawes, 1982; Lyons & Woods, 1991; Matheny, Aycock, Pugh, Curlette, & Silva, 1986; Miller & Berman, 1983). Pooling estimates across these meta-analyses suggest that about 20% of the total psychotherapy effect could indeed be attributed to nonspecific treatment components (i.e., placebo effect; $d = .18$). However, the remaining 80% of the total effect can be attributed to the unique treatment components of psychotherapy ($d = .68$). Thus, the total psychotherapy effect appears to be a combination of specific and nonspecific treatment effects ($d_{\text{Total}} = .68 + .14 = .82$).

8.4.4 Meta-Analysis and the Principle of Empirical Interpolation and Extrapolation

Empirical interpolation and extrapolation are most closely linked to the third generalizability question, in which inferences about novel universes are desired. Because individual studies are often limited in the range of u, t, o, s that are included, combining different studies may be of great benefit if each study investigates a different level of a quantitative dimension.

It is common in meta-analyses to model dose-response relationships where different studies contribute effect estimates at different dosages. Similarly, meta-analyses frequently examine the stability of treatment effects over time by combining data from studies in which effects were assessment at different time intervals after treatment ended. Such models may then be used to interpolate or extrapolate effects at levels not studied.

Recently, Shadish, Matt, Navarro, and Phillips (2000) applied an extrapolation strategy in a meta-analysis of psychotherapy outcomes in "the lab" (i.e., efficacy conditions) versus "the clinic" (i.e., effectiveness conditions). Briefly put, the "lab vs. clinic" debate arose because the vast majority of psychotherapy outcome studies are conducted in ways that are not very representative of the conditions under which therapy is actually conducted by practicing therapists. For example, lab studies often are conducted at universities with clinically inexperienced graduate student therapists who are trained intensively and specifically in a single treatment that is then applied uniformly to a highly selected patient population with a narrow range of problems that the treatment is deliberately designed to help. Such therapy is quite different from the real world of clinic therapy in which experienced therapists work in busy clinics giving an eclectic array of therapy to patients with diverse problems. If this criticism is true, then there might be serious doubts about whether the results of psychotherapy meta-analyses generalize to clinically representative conditions. Indeed, in a preliminary examination of this issue limited to child psychotherapy, Weisz, Weiss, and Donenberg (1992) concluded that the very few studies of clinic therapy that they could locate showed little or no effects compared to no treatment control. Shadish et al. (1997) revisited the same issue, asking the authors of published meta-analyses to identify studies in their data bases that were conducted outside of research labs. They found that the effects of studies approximating clinical practice were about the same as those conducted in research labs. However, like Weisz et al. (1992), Shadish et al. (1997) found very few studies of clinic therapy, and concluded that the generalizability of psychotherapy meta-analysis results to clinically representative settings was a topic that still needed far more study.

Recently, Shadish et al. (2000) reinvestigated this issue and conducted a new meta-analysis of 90 psychotherapy outcome studies, differing in the degree to which they approximate prototypical conditions of clinical practice. They concluded that therapy effects do not deteriorate over the range of clinical representativeness that was present in the 90 outcomes studies. Shadish et al. (2000) also found that effects increase with larger dose, and when outcome measures are specific to treatment. Thus, clinic therapies may be able to produce larger effects by providing longer and more intensive treatments. Moreover, some clinically representative studies used self-selected treatment clients who were more distressed than available controls, and these quasi-experiments underestimated therapy effects. Given the range of clinical representativeness in existing outcome studies, Shadish et al. (2000) extrapolated effects of an ideal study of clinically representative therapy. This projection suggests that effects are similar to those reported in past meta-analyses of studies conducted in research settings.

8.4.5 Meta-Analysis and the Principle of Causal Explanation

There are two major strategies with which causal explanation may strengthen generalized inferences based on meta-analysis. The first strategy involves decomposing domains to isolate those components that are involved in the generation and moderation of a treatment effect. Meta-analyses contribute to causal explanation to the extent that the studies allow meaningful decomposition of treatments, outcomes, persons, or settings. For instance, decomposing treatment effects in specific and nonspecific components contributes to the causal explanation of how psychotherapy effects come about. Similarly, differentiating between different settings in which services are delivered (i.e. clinic vs. lab), between levels of therapist experience and training, between types and modalities of psychotherapy, or between effects on target behaviors and peripheral symptoms contribute to a better understanding of the causal effects of the intervention.

A second strategy involves identifying the causal mediating processes that are set in motion by a treatment. Although there is nothing in theory that would prevent meta-analyses to strengthen inferences about causal mediating processes, such contributions are unlikely to be made routinely. The reason for this is that our best explanatory models are typically phrased at levels far below that of a meta-analysis aggregating across studies. For psychotherapy this may involve theories of behavior change at the level of an individual person or family. For meta-analyses to synthesize studies of micro-mediating processes at that level, theories must have reached a level of maturity and consensus such that multiple studies of the same theory and of the same micro-mediating processes are available. In many disciplines within the social and behavioral sciences, researchers tend to pursue different causal explanations and measure different explanatory processes, providing little opportunity to accumulate multiple studies from different research groups.

There are some more mundane constraints for meta-analyses of micro-mediating processes. There is the paucity of detail about treatment components in many publications. Unless additional detail can be obtained by contacting the original researchers, a meta-analysis can often contribute very little to further causal explanation at the level of micro-mediating processes. There is also a considerable amount of selective reporting that takes place when publishing study findings. Thus, a study may highlight the significant contribution of some micro-mediating processes but fail to report the nonsignificant contribution of others. Such a reporting bias (Matt & Cook, 1994) favors Type I errors and limits generalizability of meta-analytic conclusions. Finally, reported methods often reflect what was planned rather than what was achieved, in which case meta-analysts are in a poor position to accurately describe the conditions under which a particular process was observed. There are a few notable exceptions in behavioral science meta-analyses (e.g., Harris & Rosenthal, 1984; Becker, 1992; Devine, 1992; Shadish, 1992) and there may be more in the natural sciences where theories have reached more mature levels.

8.5 CONDITIONS THAT FACILITATE GENERALIZED CAUSAL INFERENCES

Generalizations based on single studies are usually weak. This is the case because an individual study can only do so much to make heterogeneous the substantive irrelevancies and explore potentially interacting moderator conditions. At the same time, the fact that many studies have been conducted on a particular topic does not guarantee valid empirical generalizations. What follows is an outline of some of the conditions that promote generalizable causal inferences.

8.5.1 Individual Programs of Research

Programs of research are collections of multiple connected studies conducted by one or more researchers or research groups on a particular research question or family of questions. Studies conducted as part of a research program play a crucial role in generating generalizable knowledge, particularly if they involve different research designs, different populations, interventions, measures, and settings. Multiple studies generated by such research programs facilitate the generalization of findings by probing the robustness of causal relationships in the face of substantively irrelevant aspects, by identifying the substantive factors that moderate an effect, and by elaborating explanatory processes. They provide evidence based on which interpolations and extrapolations can be based and the proximal similarity between sample and universes can be justified.

8.5.2 Integrative Reviews

Different from individual programs of research, integrative reviews involve collections of primary studies that are not necessarily well connected or orchestrated by a network of researchers. Instead, such studies often cover many different research programs, published over many decades, by researchers from different countries on different continents, subscribing to different research paradigms. While integrative reviews are less likely to advance causal explanations – a particular strength of individual research programs – their major contribution is likely to come from exploring heterogeneous irrelevancies across the large diversity of populations, outcomes, interventions, historical and cultural contexts, and so forth. Because integrative reviews are likely to involve large numbers and diverse characteristics, they provide a particularly good opportunity to explore the robustness of a causal relationship, to investigate factors that may moderate its direction or size, and to interpolate and extrapolate to new domains.

While meta-analyses have many potential benefits for facilitating generalized inferences, they cannot provide a panacea for generalization questions. As is the case with quasi-experimental designs of primary studies, the nonexperimental nature of meta-analyses makes it necessary to carefully examine and rule out plausible alternative explanations to claims of causal generalization (Matt & Cook, 1994).

8.5.3 Critical Multiplism

Focused research programs and integrative reviews will provide little evidence to support causal generalizations if individual studies rely on the same subject populations, recruitment strategies, interventions, measures, and settings. Such shared research design and interventions characteristics are likely to produce the same method biases masquerading and confounding underlying causal effects. Synthesizing individual studies that share the same biases yields biased meta-analytic findings, leaving the meta-analytic evidence of little use to support any of the principles discussed above.

To provide a rich foundation for causal generalization, critical multiplist methods should be applied whenever possible at the level of the individual study, the focused research program, or the integrative review (Cook, 1982; Shadish, 1993). To mention a few, such methods call for the investigation of multiple methods to assess outcome, multiple settings to investigate interventions, several smaller studies rather than a single large study, multiple subject population, multiple research groups, and so forth.

8.5.4 Public Debates

By their nature, causal generalizations are at the center of many public policy debates. Such debates play a crucial role in forcing out hidden assumptions and assuring that important stakeholders have been taken into account.

Because of public debates, new sources of bias or important new contextual conditions may be discovered. Public debate will draw attention to the personal and public costs and benefits of more stringent or liberal policies. In addition, public debates help establish the extent to which more liberal or more conservative generalizations may be implemented. In the case of California's public policy regarding second-hand smoke exposure, public debates (including a public referendum) led to the adoption of a more liberal generalization regarding the health effects of low-level second-hand smoke exposure. In this instance, the public health benefits of "overgeneralizing" (i.e., second-hand smoke exposure is toxic at any level) were found to outweigh the risk of policies mandating ventilation systems to reduce second-hand smoke exposure levels even though such harsher policies interfere with the civil rights of smokers.

8.6 CONCLUSIONS

Rarely – if ever – do researchers and consumers of research believe that findings apply only to the specific circumstances of a specific study. Instead, we believe – or behave as if – findings apply to larger domains of persons, interventions, outcomes, settings, and times than were included in a study. At issue are three types of empirical generalizations with respect to persons, treatments, outcome, settings, and times. The first type of generalization involves inferences to target universe based on specific samples, a form of inductive empirical generalizations. The second involves inferences across target universes or across sub-universes, a form of deductive empirical generalizations. The third involves generalized inferences to novel universes, closely related to empirical interpolation and extrapolation.

Empirical generalizations are best achieved through the synthesis of multiple studies, conducted by many research teams, with different populations, in different settings, with multiple operationalizations of interventions and outcomes. One form of research synthesis, meta-analysis, has particularly great promise to facilitate generalized inferences. Meta-analysis provides no shortcuts or guarantees for generalizations. However, it does provide research design and analytical tools to conduct principled investigations of generalizability claims and increases the likelihood of better inferences compared to individual studies.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum.

Babbie, E. R. (1995). *The practice of social research* (7th ed.). Belmont: Wadsworth.

Becker, B. (1992). Models of science achievement: Forces affecting male and female performance in school science. In T. D. Cook, H. Cooper, D. S. Corday, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A case-book* (pp. 209–281). New York: Russell Sage Foundation.

Bhaskar, R. (1978). *A realist theory of science* (2nd ed.). Atlantic Highlands, NJ: Humanities Press.

Bowers, T. G., & Clum, G. A. (1988). Relative contribution of specific and nonspecific treatment effects: Meta-analysis of placebo-controlled behavior therapy research. *Psychological Bulletin*, *103*, 315–323.

Brunswik, E. (1947). *Systematic and representative design of psychological experiments.* Berkeley, CA: University of California Press.

Brunswik, E. (1952). *The conceptual framework of psychology.* (Int. Encycl. unified Science, v. 1, no. 10.). Chicago, IL: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Campbell, D. T. (1969). Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin*, *98*, 388–400.

Clum, G. A., Clum, G. A., & Surls, R. (1993). A meta-analysis of treatments for panic disorder. *Journal of Consulting and Clinical Psychology*, *61*, 317–326.

Cook, T. D. (1982). Postpositivist critical multiplism. In L. Shotland & M. M. Marks (Eds.), *Social science and social policy.* Newbury Park, CA: Sage.

Cook, T. D. (1990). The generalization of causal connections. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data.* (DHHS Pub. No. 90–3454). Washington, DC: U.S. Department of Health and Human Services.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them* (pp. 39–82). (New Directions in Program Evaluation, No. 57). San Francisco: Jossey Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Cormier, S. M., & Hagman, J. D. (1987). *Transfer of learning: Contemporary research and applifications.* San Diego: Academic Press.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Fransisco: Jossey Bass.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *British Journal of Statistical Psychology*, *16*, 137–163.

D'Amato, R. J., Loughnan, M. S., Flynn, E., & Folkman, J. (1994). Thalidomide is an inhibitor of angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 4082–4085.

Devine, E. C. (1992). Effects of psychoeducational care with adult surgical patients: A theory-probing meta-analysis of invention studies. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A case-book* (pp. 35–82). New York: Russell Sage Foundation.

Estes, W. K. (1978). *Handbook of learning and cognitive processes.* Hillsdale, NJ: Lawrence Erlbaum.

Groeben, N., & Westmeyer, H. (1975). *Kriterien psychologischer Forschung* [Criteria for psychological research]. München: Juventa.

Hammond, K. R. (1948). Subject and object samling – A note. *Psychological Bulletin*, *45*, 530–533.

Hammond, K. R. (1951). Relativity and representativeness. *Philosophy of Science*, *18*, 208–211.

Harris, M. J., & Rosenthal, R. (1984). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*, 363–386.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Hommel, B., & Prinz, W. (1997). *Theoretical issues in stimulus–response compatibility.* New York: Elsevier.

Kazdin, A. E., Bass, D., Ayers, W. A., & Rodgers, A. (1990). Empirical and clinical focus of child and adolescent psychotherapy research. *Journal of Consulting and Clinical Psychology*, *58*, 729–740.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome. Smith and Glass' conclusions stand up under scrutiny. *American Psychologist*, *37*, 504–516.

Lenz, W. (1962). Thalidomide and congenital abnormalities. *The Lancet (Volume 1)*, 45.

Lyons, L. C., & Woods, P. J. (1991). The efficacy of rational emotive therapy: A quantitative review of the outcome research. *Clinical Psychology Review*, *11*, 357–369.

Matheny, K. B., Aycock, D. W., Pugh, J. L., Curlette, W. L., & Silva, K. A. (1986). Stress coping: A qualitative and quantitative synthesis with implications for treatment. *Counseling Psychologist*, *14*, 499–549.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 503-520). New York: Russell Sage Foundation.

Matt, G. E., & Navarro, A. M. (1997). What meta-analyses have and have not taught us about psychotherapy effects: A review and future directions. *Clinical Psychology Review*, *17*, 1-32.

McBride, W. G. (1961). Thalidomide and congenital abnormalities. *The Lancet (Volume 2)*, 1358.

Miller, R. C., & Berman, J. S. (1983). The efficacy of cognitive behavior therapies: A quantitative review of the research evidence. *Psychological Bulletin*, *94*, 39–53.

National Academy of Sciences. (1997). *Technology transfer systems in the United States and Germany: Lessons and perspectives.* Washington, DC: National Academic Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Pfeiffer, R. A., & Kosenow, W. (1962). Thalidomide and congenital abnormalities. *The Lancet (Volume 1)*, 45–46.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach.* Oxford: Clarendon Press.

Shadish, W. R. (1992). Do family and marital therapy change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. Cooper, D. S. Corday, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A case-book* (pp. 129–208). New York: Russell Sage Foundation.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. In L. Sechrest (Ed.), *Program evaluation: A pluralistic enterprise* (pp. 13–57). (New Directions in Program Evaluation, No. 60). San Francisco: Jossey-Bass.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*, 512–529.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, S., Crits-Christoph, P., Hazelrigg, M., Jorm, A., Lyons, L. S., Nietzel, M. T., Thompson Prout, H., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). The effects of psychotherapy conducted under clinically representative conditions. *Journal of Consulting and Clinical Psychology*, *65*, 355–365.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: The John Hopkins University Press.

U.S. Food and Drug Administration. (1998). *The Center for Drug Evaluation and Research (CDER) Handbook.* Washington, DC: Food and Drug Administration.

Verheul, H. M., Panigrahy, D., Yuan, J., & D'Amato, R. J. (1999). Combination oral antiangiogenic therapy with thalidomide and sulindac inhibits tumour growth in rabbits. *British Journal of Cancer*, *79*, 114–118.

Webster. (1986). *Webster's third new international dictionary of the English language unabridged.* Springfield, MA: Merriam-Webster.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic. Effects of child and adolescent psychotherapy. *American Psychologist*, *47*, 1578–1585.

Zadeh, L. A., & Yager, R. R. (1987). *Fuzzy sets and applications: Selected papers.* New York: Wiley.

9

Meta-Analysis – Not Just Research Synthesis!

Uwe Czienskowski

Max Planck Institute for Human Development
Berlin

Summary

Meta-analysis represents an advanced methodological approach to the (quantitative) synthesis of different studies within a research field. However, meta-analytical integration is mostly not pursued further after several moderators have been identified that are responsible for much of the heterogeneity of results across primary research. In this chapter, the necessity of completing a meta-analytical integration of previous research by independently conducting primary research is stressed. It is shown that this approach to meta-analysis allows one to distinguish between merely potential moderators and real ones. This approach particularly considers meta-analysis a tool for the generation of new hypotheses as well as for the design of precise and sensitive decision studies. As an example, research on the self-reference memory effect is presented to demonstrate how to use meta-analysis not only to integrate a research field, but also to identify theoretical and empirical shortcomings within primary research. Discussing several possible objections against meta-analysis, it is concluded that meta-analysis, if adequately conducted and interpreted, is not only a tool for research integration, but may also be used in a theoretically fruitful way.

9.1 INTRODUCTION

During the last decades, meta-analysis has frequently been proven to be a collection of useful statistical techniques for the quantitative integration of results from different fields (e.g., Cooper & Hedges, 1994; Hedges, 1992). In comparison with other approaches to assessing the state of research in a specific area (e.g., narrative review, simple vote counting; see Bushman, 1994) meta-analysis relies upon statistical indices that represent the magnitude of an empirical effect, investigated by means of, for example, experimental or correlational techniques on a common scale of measurement that is independent of a study's sample size. Using the notion of this so-called *effect size*(ES; e.g., Richardson, 1996; Rosenthal, 1994; Tatsuoka, 1993), different meta-analytical approaches have been developed, depending not only on the type of ES but also on the kind of investigation predominantly used within a research field. For example, collections of studies using Cohen's *d* (standardized difference of means; see Cohen, 1988) as ES for experiments or quasi-experiments may be integrated using procedures described by Hedges and Olkin (1985), whereas research described best by variance-compound-directed ESs, for example, estimated ω^2 (Hays, 1994) or the correlation coefficient *r*, might be integrated by applying a "psychometric" meta-analysis as described by Hunter and Schmidt (1990; see also Johnson, Mullen, & Salas, 1995; Schmidt & Hunter, 1999). Furthermore, different procedures for the combination of ESs for categorical data, for example, *rate ratios* or *odds ratios* (see Fleiss, 1994), are widely used in medicine and epidemiology (e.g., Petitti, 1994). In general, meta-analytical integration is directed to present an average ES for a group of studies investigating the same empirical effect. In the simplest case, the mean ES for *i* studies to be integrated can be computed as a sum of *i* weighted ESs, divided by the sum of *i* weights (Shadish & Haddock, 1994). It is generally assumed that there are only two possible sources of variation of ESs: Variation can occur by chance if all studies share a common population ES, and additional systematic variation between studies can arise if they do not. In the latter case, categorical variables (moderators) are investigated to determine if they are responsible for this systematic variation. This is known as the moderator analysis approach. Another strategy to cope with unexpected systematic variation is known as the random effects model. This statistical model does not assume that each study effect estimates the same population effect, but rather that each single effect represents a random variable with its own distribution (e.g., Raudenbush, 1994; Shadish & Haddock, 1994).

In this chapter, the moderator analysis approach will be discussed in more detail. It will be shown that an independent investigation of moderators is necessary to cope with uncertainty of the state of potential moderators, especially if data from experimental studies have been integrated meta-analytically. The usefulness of directly manipulating moderating variables in subsequent experimentation has already been demonstrated, for example, in a study conducted by Bornstein, Kale, and Cornell (1990; see also Eagly & Wood, 1994). However, although this study has been inspired by a previous meta-analysis (Bornstein, 1989), it is not directed to an experimental evaluation of moderators that have been drawn from meta-analytical integration. Furthermore, the use of research syntheses for theoretical progress has been discussed extensively by Cook et al. (1992) as well as Miller and Pollock (1994).

In this chapter, it will be demonstrated how information on a tentatively hypothesized moderating variable can be used for a direct evaluation of its actual meaning. In addition, it will be shown that by tying meta-analysis to primary experimental research, more general problems of meta-analytic approaches can be solved in a simple way. Most important, it will be demonstrated how meta-analysis as well as primary experimental research inherit the specific advantages of each other by this link, and how this link may lead to theoretical progress that cannot be obtained without the interplay of meta-analytical integration and experimental validation.

9.2 MODERATORS IN RESEARCH INTEGRATION: AN EXAMPLE

Suppose we conduct a meta-analysis on a specific memory effect that has been investigated in, say, 72 different experiments. Most of these experiments support the idea that presenting an orientation task like "Does the following word describe you?" leads to better recall for subsequently presented words than the orientation task "Does the following word describe Bill Clinton?". This recall difference is known as the so-called "self-reference effect" in memory (SRE). Suppose further that our meta-analysis supports the conclusion that the first condition (self-reference) actually does result in better memory performance than the second (other-reference). The average ES for this comparison is about $r = .25$ (we conducted a meta-analysis following the approach of Hunter & Schmidt, 1990). Our analysis reveals noteworthy heterogeneity, so that a moderator analysis seems to be indispensable. Fortunately, we are able to identify two variables, hardly compared within single studies but quite often between studies: intimacy with the person referred to in the other-reference condition (*high* vs. *low*) and type of material presented (*adjectives* vs. *nouns*). Further analysis has revealed that both variables seem to moderate our previously noted results. The magnitude of recall enhancement under a self-referential instruction is only marginal when compared with a high-intimacy target person in the other-reference condition, but substantial if a low-intimacy person is referred to. However, this effect is only observable for adjectives; it disappears when nouns are to be recalled. To sum up, on a meta-analytical level we observed a pattern indicating an interaction between intimacy and word type (see Figure 9.1).

9.3 WHAT IS THE REAL MEANING OF A MODERATOR?

The question remains, however, whether a difference between two or more groups of studies that has been identified by means of a moderator analysis

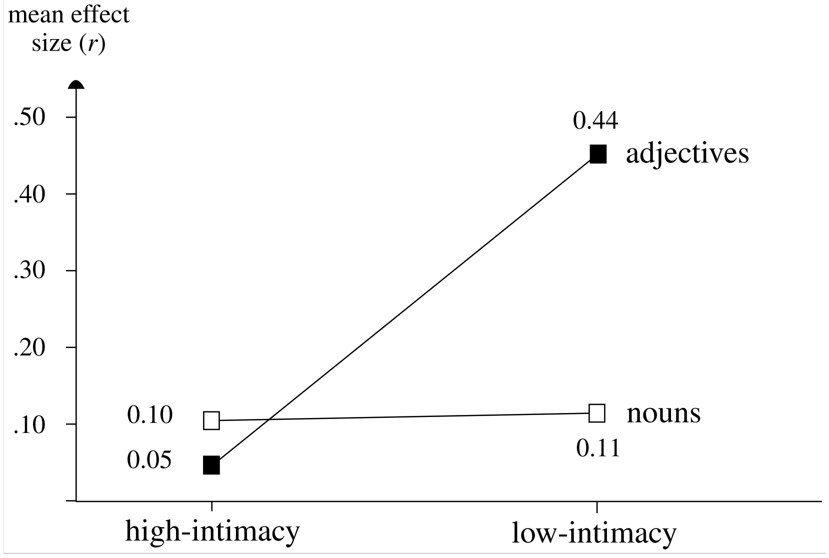


Figure 9.1 Interaction of SRE for intimacy and word type.

actually does represent a valid difference, or whether it merely represents an explanation by chance. This uncertainty can be reduced by selecting a more appropriate statistical model; that is, to interpret an effect size as a random effect allows a higher degree of generalizability than to consider it as fixed. Still, the basic problem remains unsolved even if we regard an effect size as random: A successful categorization of a collection of effect sizes by a specific variable that differentiates *between* different studies does only allow a post-hoc explanation of some of the variation between effect sizes that differs from random error. But in this case a meta-analytical approach is basically correlational (see Hall & Rosenthal, 1991); that is, no causal relationship can be established with this procedure. Even if a moderating variable can differentiate sufficiently between subgroups of effect sizes, the conclusion cannot be drawn that this variable has actually caused these differences. Since causation can generally be inferred only if based on experimental manipulations with results supporting this relation, a strong requirement can be formulated concerning the state of a moderator: A moderating effect of a variable that has been identified post-hoc based on meta-analytical results should be treated as a tentatively accepted potential moderator. Its state as a real moderator has to be evaluated by means of independent follow-up experiments. If this validation procedure is omitted, a scientific explanation of the differences between primary studies by means of the supposed moderator under study is not justified, even if a statistical explanation of the observed heterogeneity has been obtained by meta-analysis.

9.4 TESTING MODERATOR HYPOTHESES EMPIRICALLY

At this point we can pick up the thread again and discuss in more detail how to proceed with the result pattern presented in Figure 9.1. It has been noted that the meta-analysis on the difference in memory performance between recalling words under a self-referent task and an other-referent task seems to be moderated by at least two moderating variables, intimacy and word type. Taking both as potential moderators as discussed above, we may now use the meta-analytical results directly. The average ES for a group of integrated studies actually represents the most exact estimate of a population effect available, since it covers more data than a single study could provide. Furthermore, we are not required to conduct an exploratory study to estimate an expected effect. Relying on the given ES estimate, we are now able to design a decision study that enables us to decide with maximum precision whether our potential moderator does actually have explanatory power for the observed heterogeneity of ESs or whether the meta-analytical results should be taken merely as chance hits. Let us consider the result that the memory advantage of a self-referential orientation task compared with low-intimacy other-referential encoding is strong for adjectives ($r = .44$), but only marginal for nouns ($r = .11$). For the moment we will put aside a discussion of this puzzling result and the interesting question of whether the contrast between adjectives and nouns is meaningful at all. We will return to these topics below.

If we are interested in testing for the above-mentioned result pattern, we only have to specify the smallest difference to be detected between two conditions that should be shown as different and to specify appropriate levels for the first-order and second-order errors (e.g., .05 and .20). Then the required sample size for an adequately designed experiment (e.g., Keppel, 1991) can be computed or taken by power tables as published by Cohen (1988). In the case at issue, a group size of about 22 persons is required to test for the difference of $r = .39$ between two SREs (low- and high-intimacy others) for adjectives both exact and sufficiently sensitive ($\alpha = .05, \beta = .20$). If only the SRE with low-intimacy other persons ($r = .44$) is to be tested, $N = 18$ are required, provided that the references will be compared within subjects (as is mostly done in primary research).

As the reader may already have guessed, our examples are not fictitious but are results drawn from a meta-analysis on the self-reference effect that has actually been published (Czienskowski, 1997). In comparison with the meta-analysis on the SRE by Symons and Johnson (1997), the main aim of the integration by Czienskowski was to identify subordinate moderating patterns, using a hierarchical breakdown strategy (Hunter & Schmidt, 1990). Furthermore, it was attempted to test some predictions generated by this meta-analysis by means of further experiments. For example, Czienskowski (1997, 1998) reports experiments showing that intimacy with a person referred to seems actually to be a central factor that determines the magnitude of the investigated advantage of self-referential encoding. If an extremely low-intimacy other-reference condition is used, Czienskowski (1998, Exp. 2) reports a much stronger SRE

(r = .61). Further experiments indicate that for average low-intimacy other-referents the SRE is approximately the same as predicted from meta-analysis (r = .43 resp. r = .41), but no remarkable difference can be obtained for the contrast between self-reference and high-intimacy other-reference (Czienskowski, 1997, Exp. 2; 1998, Exp. 1). Moreover, other results seem to confirm the prediction that only the use of adjectives, but not nouns, can produce the difference reported above. Problems related to the investigation of an SRE using nouns will be discussed below.

To sum up, it seems very promising to take an apparent moderator from meta-analysis as a merely potential moderator, which has to be tested independently. Since an empirical test of a prediction generated by moderator findings may also fail, this requirement is not at all trivial but indispensable to protect a meta-analysis against the obvious problem of "capitalizing on chance".

9.5 IS META-ANALYSIS USEFUL FOR THEORY DEVELOPMENT?

Although this leading question is answered affirmatively by most researchers who rely on meta-analytical techniques (e.g., Hall, Rosenthal, Tickle-Degnen, & Mosteller, 1994; Cook et al., 1992), it is often taken for granted that meta-analysis represents a powerful statistical toolbox that can be used to integrate different studies but that has no influence on the development of scientific theories. Actually, an ignorant use of meta-analytic tools may result in incorrect conclusions and remarkable confusion. But this is not only true for meta-analysis but for all advanced statistical technologies. To show how an adequate meta-analytical approach may in fact be used in a theoretically fruitful way, I will now focus on a special problem of the meta-analytical results referred to above.

As previously noted, it seems puzzling that a strong SRE is obtained for a low-intimacy other-referent condition if adjectives are used, but not if nouns are used. A closer look at the meta-analytical database suggests that these results actually may be an artifact. The database of studies included in the meta-analysis does not contain any study that compares adjectives with abstract nouns, but only with rather concrete nouns. On the other hand, one study that investigates a somewhat different kind of SRE (i.e., the recall difference between a self-reference and a merely semantically directed orientation task) using at least partly abstract nouns (Bock, 1986), indicated a very strong SRE. Since especially the distinction between concreteness and abstractness represents a central dimension for the explanation of memory performance (e.g., Gee, Nelson, & Krawczyk, 1999; Holcomb, Kounios, Anderson, & West, 1999; Marschark & Surian, 1992; Paivio, Walsh, & Bons, 1994), it can reasonably be assumed that the meta-analytical results could be confounded with a further but uncontrolled factor called "concreteness of word type". To examine this assumption further, Czienskowski (1997) reports an experiment that compares self-reference and low-intimacy other-reference using adjectives and matched abstract nouns. For both materials a strong SRE has been found, which is completely incompatible with results yielded by the meta-analysis. In a second experiment, the assumption was tested that the SRE is obtained if abstract nouns but not concrete ones are used, and moreover, that this effect holds only for low-intimacy and not for high-intimacy other-referents. Planned simple effects analyses and simple comparisons are reported that support the expected result pattern. Simple comparisons between low-intimacy other-reference on the one hand and high-intimacy others and self on the other hand indicate strong memory differences (about $r = .41$) only for abstract nouns, whereas concrete nouns do not produce any detectable difference. More important might be that this effect is only due to a reduced recall performance for low-intimacy other-referents under the abstract noun condition. When exclusively explaining the SRE by referring to special features of self-referent encoding processes (e.g., as Rogers, 1981, does), this result still remains puzzling, because then it cannot be explained why an SRE should only occur when using abstract nouns. Moreover, the results reported seem to indicate that the SRE, at least for the comparison of self and others, is not an effect of enhanced self-referent encoding but of reduced recall for abstract material if low-intimacy others are referred to.

Focussing now on the fact that the critical difference causing the result pattern just discussed seems to be the concreteness or abstractness of the stimulus word presented, a reasonable assumption might be that the so-called SRE is actually a subordinate effect occurring only if other conditions are absent that could support encoding processes. More precisely, in the above case a concreteness effect as described, for example, by the dual-coding theory (DCT; Paivio, 1971, 1991) seems to be superior, whereas a self-reference effect (which rather seems to be an effect of intimacy or familiarity) takes place only if the concreteness of a stimulus is too low for promoting memory encoding. Czienskowski and Giljohann (2002) report two experiments indicating that a recall advantage of self-reference and high-intimacy other-reference can only be detected when abstract nouns are presented. With concrete nouns, the recall under self-reference is substantially lower (about $r = .30$) than for both other-reference conditions. The results confirm the expectation that only the absence of a possibility to encode information pictorially may result in a strong and unequivocal self-reference effect and support the view that the self-reference effect is not a general memory effect, but only a subordinate one. It may buffer the decrease of memory performance if pictorial coding is not possible, but it is not able to compensate this.

9.6 META-ANALYSIS AS A TOOL: IDENTIFYING THEORETICAL DEFICIENCIES AND NEW HYPOTHESES

Collecting the evidence from different sources, we are now able to conclude that an adequate application of meta-analysis in a rather more developed field of empirical research does not necessarily represent mere integration but can also be used for the theoretical refinement or even reformulation of existing and frequently tested theoretical assumptions. In the case discussed above, the quantitative integration of studies investigating a rather prominent effect of cognitive social psychology has revealed both theoretical problems as well as new empirical hypotheses that might not have been detected without the application of meta-analytical methods. Admittedly, we can imagine a situation in which these problems and new hypotheses might have been developed by an attentive researcher interested in the SRE and aware of the DCT without making use of meta-analysis. However, in this case the predictions and the tests conducted subsequently would be much more imprecise than the expectations generated by research integration because only rough predictions of expected effects are possible. Thus, the state of the obtained results would remain rather unclear. Moreover, the evidence for a primary result is inferior compared to the evidence given by a decision study based on meta-analytical predictions drawn from a rather large set of primary studies.

It is quite obvious that this approach is not only applicable to the quantitative integration of research, but also, for example, if more evidence is required for an appropriate use of a therapy, for the development of educational strategies, or for a decision between two competing theories of memory. On the contrary, different fields of primary research could profit from this approach. In general, the determination of an adequate sample size for an experiment requires fixing an ES as precisely as possible if it is to be tested adequately by an experiment (i.e., both controlling the risk of rejecting both types of hypotheses, i.e., H_0 and H_1, falsely). Since theories in behavioral sciences are mostly not able to specify the exact magnitude of an effect to be tested, the integration of meta-analytical procedures into the process of designing primary studies should be seen as a opportunity to avoid wasting effort on conducting uneconomical (i.e., the sample size is too large for an effect to be tested) or meaningless (i.e., the sample size is too small for sensitively detecting an effect) experiments. The integration of meta-analytical procedures seems to be particularly favorable, too, just because several studies (e.g., Cohen, 1962; Sedlmeier & Gigerenzer, 1989) have reported that the average power of experiments published in certain journals only amounts to about .50 or even less. Thus, the use of meta-analytical tools can been seen as an indispensable supplement to the use of other design tools as, for example, power analysis, at least in a field of research that is rather developed.

9.7 CONCLUSION

In the previous parts of this chapter, a perspective on meta-analysis has been developed that is motivated mainly by requirements predominantly stated within the realm of primary experimental research. Hence, the orientation is directed to fields that can be investigated experimentally, at least in principle. With this restriction in mind, I can now discuss some conclusions that could bring about a new evaluation of several objections directed against meta-analytical approaches. I will concentrate on some topics that are generally referred to as main problems of meta-analysis (e.g., Glass et al., 1981; Beelmann & Bliesener, 1994).

Let us first investigate the so-called apples-and-oranges problem. It is stated that, because the main feature of a study is its theoretical background, different operationalizations indicate different concepts so that different studies cannot be compared. But this conjecture is not convincing, since it implies the impossibility of modifying theories with the help of statistical meta-analysis. However, theories are quite often affected or even falsified by data. Since an empirical hypothesis cannot be rejected by any a priori argument, no a priori evaluation of the empirical relevance of any potential moderator hypothesis is possible. If a moderator variable can be identified that is able to explain the difference between studies not merely statistically, but also within an independent study, a theory not predicting this effect must be characterized as deficient (for sure, many possible moderators are theoretically irrelevant or even trivial, but since a meta-analyst is expected to be an expert in the field to be integrated, these irrelevant "moderators" will probably be sorted out early). Thus, the conjecture that meta-analysis is not able to provide more theoretical information than primary research can be refuted. Moreover, from a primary research point of view it could actually be advantageous to analyze some theoretical relations across different studies before more primary research is conducted. A meta-analytic integration of research and its use for the design of new studies does allow a goal-directed and precise search for theoretical relations to be identified empirically. By comparison, without any information from research synthesis, tests of these effects would be imprecise at best, but mostly these effects would not be identifiable at all.

A more general problem of meta-analysis could be its epistemological assumption of the possibility of accumulating scientific knowledge. From the author's point of view, the method of meta-analysis is indifferent to the problem if accumulation within the progress of science is really possible. Actually, it does not seem very useful to integrate empirical results of studies from different research fields, even if the results seem to be similar or at least comparable to each other. However, within a field in which a specific research question is investigated, a meta-analytical integration across different studies may be used to acquire higher precision of empirical statements. If meta-analysis is thus simply taken as a statistical tool, basically neither better nor worse than other statistical tools widely used (or misused) in behavioral sciences, there is no need to emphasize some of its problems more than, say, the problem of applying analysis of variance on ordinal dependent variables. Meta-analysis can be misused as much as other statistical procedures can, but there are many research problems that can be solved by this approach, therefore rejecting the importance of meta-analytical research integration would be throwing the baby out with the bathwater.

Some further problems of meta-analytic approaches, for example the "garbage in – garbage out problem" (i.e., a meta-analysis that integrates across poorly designed studies will probably be biased), can be resolved in a straightforward manner by explicitly testing potential moderators. The crucial question is whether the relevance of a possible moderator can be confirmed empirically. An empirical test will yield significant differences between different levels of the moderator if the effect size estimation from meta-analysis does actually represent a population effect, but it will fail if no population difference exists. This means that if moderators are indicated on the basis of biased samples or simply by chance, these potential moderators will be rejected if their test fails, otherwise the effect seems to be real, even if indicated by poorly designed studies.

Finally, to sum up the arguments given in this chapter, we can state that meta-analytical methods are powerful techniques and should be seriously considered if their integration into the methodological toolbox, especially the toolbox of primary research, could have advantages. On the other hand, research synthesis could benefit from this link, too, because some severe shortcomings that could affect the meaning of a research synthesis can be avoided if the presented techniques are applied. This approach seems especially useful if the real meaning of moderating variables is to be understood.

REFERENCES

Beelmann, A., & Bliesener, T. (1994). Aktuelle Probleme und Strategien der Meta-Analyse [Recent problems and strategies of meta-analysis]. *Psychologische Rundschau*, *45*, 211–233.

Bock, M. (1986). The influence of emotional meaning on the recall of words processed for form or self-reference. *Psychological Research*, *48*, 107–112.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Bornstein, R. F., Kale, A. R., & Cornell, K. R. (1990). Boredom as a limiting condition on the mere exposure effect. *Journal of Personality and Social Psychology*, *58*, 791–800.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193–213). New York: Russell Sage Foundation.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook.* New York: Russel Sage Foundation.

Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3–14). New York: Russell Sage Foundation.

Czienskowski, U. (1997). Selbstbezug: Eine besonders wirksame Enkodierungsstrategie? Meta-Analyse und experimentelle Moderatorprüfung [Self-referencing: A very effective encoding strategy? Meta-analysis and experimental validation of moderators]. *Zeitschrift für Experimentelle Psychologie*, *44*, 361–393.

Czienskowski, U. (1998). Kann experimentelle psychologische Forschung von Meta-Analysen profitieren? [Can experimental psychological research benefit from meta-analyses?]. In K. C. Klauer & H. Westmeyer (Eds.), *Psychologische Methoden und Soziale Prozesse* (pp. 210–229). Lengerich: Pabst Science Publishers.

Czienskowski, U., & Giljohann, S. (2002). Intimacy, concreteness, and the 'self-reference effect'. *Experimental Psychology*, *49*, 73–79.

Eagly, A. H., & Wood, W. (1994). Using research syntheses to plan future research. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 485–500). New York: Russell Sage Foundation.

Fleiss, J. L. (1994). Measures of effect size for categorial data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Gee, N. R., Nelson, D. L., & Krawczyk, D. (1999). Is the concreteness effect a result of underlying network interconnectivity? *Journal of Memory and Language*, *40*, 479–497.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis. Issues and methods. *Communication Monographs*, *58*, 437–448.

Hall, J. A., Rosenthal, R., Tickle-Degnen, L., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 17–28). New York: Russell Sage Foundation.

Hays, W. L. (1994). *Statistics for psychologists.* New York: Holt, Rinehart & Winston.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, *17*, 279–194.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Holcomb, P. J., Kounios, J., Anderson, J. E., & West, W. C. (1999). Dual coding, context availability and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 721–742.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three meta-analytic approaches. *Journal of Applied Psychology*, *80*, 94–106.

Keppel, G. (1991). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice-Hall.

Marschark, M., & Surian, L. (1992). Concreteness effects in free recall: The roles of imaginal and relational processing. *Memory and Cognition*, *20*, 612–620.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457–483). New York: Russell Sage Foundation.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, *45*, 255–287.

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1196–1204.

Petitti, D. (1994). *Meta-analysis, decision-analysis, and cost-effectiveness analysis.* Oxford: Oxford University Press.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers*, *28*, 12–22.

Rogers, T. B. (1981). A model of the self as an aspect of the human information processing system. In N. Cantor & J. F. Kihlstrom (Eds.), *Personality, cognition and social interaction* (pp. 193–214). Hillsdale, NJ: Lawrence Erlbaum.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, & Salas (1995). *Journal of Applied Psychology*, *84*, 144–148.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russel Sage Foundation.

Symons, S. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*, 371–294.

Tatsuoka, M. (1993). Effect size. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 461–479). Hillsdale, NJ: Lawrence Erlbaum.

Acknowledgments

I would like to thank Hubert Feger, Bärbel Knäuper, Barbara Krahé, and Mansur Lalljee for helpful comments concerning this chapter.

Part II

Applications

10

The Application of Methods of Meta-Analysis for Heterogeneity Modeling in Quality Control and Assurance

Dankmar Böhning

Working Group: Biometry and Epidemiology
Institute for International Health, Joint Center for Humanities and Health Sciences
Humboldt-University in Berlin / Free University Berlin

Uwe-Peter Dammann

Asta-Medica, Werk Künsebeck
Halle-Künsebeck

Summary

In the past few years meta-analysis has become increasingly popular in many areas of science such as medicine and pharmacy, psychology, and other social sciences. In these areas of application meta-analyses have been performed in order to obtain a pooled estimate of various single studies. Obtaining a single summary measure implicitly assumes homogeneity of these studies, that is, the results of individual studies differ only by chance. In this case a combined estimate of the individual studies provides a powerful and important result. However, this pooled estimate may be seriously misleading if study conditions are heterogenous. Thus, an approach which considers meta-analysis as a study over studies has increasingly been advocated. This approach seeks to investigate heterogeneity between studies. An important feature of this type of meta-analysis lies in the fact that it tries to identify factors which cause heterogeneity. It is the aim of this contribution, in corporation with the unit of quality assurance of ASTA Medica at location Künsebeck, to extend this approach appropriately to the area of quality control, where batches of the produced

goods replace the role of studies in medicine or the social sciences. Clearly, in this setting an investigation of heterogeneity is equally attractive, since identification and modeling of heterogeneity help to improve the production process.

10.1 INTRODUCTION AND PREVIEW

The chapter reviews an approach which enables a global perspective on aspects of homogeneity and heterogeneity which occurs in quality control and quality assurance in the pharmaceutical industry. In conventional meta-analysis, investigations are done in such a way that a specific measure can be computed utilizing numerous single studies. Frequently, statistical questions of efficiency are dominating in the literature (Hedges & Olkin, 1985). Efficiency is achieved by pooling the various single studies, thus yielding an increased sample size. This procedure, no doubt, is of great benefit, if the various studies to be combined in the meta-analysis have emerged under comparable conditions and are different in a statistical sense only by chance. This is the situation of *homogeneity*. However, pooled analysis is often considered problematic if study conditions are heterogenous, especially if the interpretation of pooled estimators is kept in a traditional way.

The chapter at hand underlines parallel aspects of meta-analysis and quality control. The cornerstone of this analogy are the numerous batches which are drawn in quality control for monitoring purposes, which play the role of the single studies in meta-analysis. Here, measures of interest are frequently count variables (counts of contamination particles) or other quality indices. In this situation – even if homogeneity conditions are present – deviations from a given standard might occur as well. It is quite important whether these deviations might have emerged from a homogenous process (as random variations) or are due to certain *heterogeneities* present in the production process. By means of the mixture distribution analysis, we are able to model potentially present heterogeneity and, further on, to classify each batch into one of the heterogeneity components. This might allow the researcher to diagnose certain common attributes and therefore enables him to explore the causes of heterogeneity.

10.2 LEGAL BACKGROUND FOR PHARMACEUTICAL PRODUCTION

Pharmaceutical production of drug products and drug substances is regulated worldwide by the rules of Good Manufacturing Practices. For Europe and Germany, producers have to follow the regulations of

- Arzneimittelgesetz (AMG)
- EU-Guideline for Good Manufacturing Practices (1989)

- "Betriebsverordnung für pharmazeutische Unternehmer" (PharmBetrV 1994)

Production and quality control of drug products and drug substances have to recognize state of the art and current worldwide practices in accordance with the application. All procedures used in production and quality control must be validated and regularly revalidated. Drug products are mainly produced in batches, which should conform with the specification from batch to batch. Drug products brought into the market should be produced and controlled according to the application and the quality has to be confirmed before a batch can be released for distribution.

The quality of a drug product or a drug substance is defined by identity, assay, chemical, physical and biological properties. A batch is the quantity of a drug produced under suitable uniform conditions to guarantee a homogeneous quality.

10.3 THE TASKS AND OBJECTIVES OF QUALITY ASSURANCE IN PHARMACEUTICAL INDUSTRY

The production of drugs is accompanied by

- batch- and product related in-process controls (on-line)
- batch- and product related controls (off-line)
- not batch and not product related controls

Parenteral drugs are products which have to comply with additional, specific properties like sterility and are essentially free of visible particles because of their parenteral application. Sterility is controlled by a sterility test which is a destructive test on limited samples of a batch. In connection with in-process controls for the clean environment of rooms, air, surface, and personnel hygiene during production, especially parenteral drugs produced by aseptic processing sterility can be assured in all parts of a batch.

Each parenteral container is controlled by a 100%-inspection for particulate matter. The quality of this inspection is controlled by samples which are again inspected for subvisual particles. These are destructive tests on a limited number of samples. The quality is evaluated on the basis of a quality index like the one which can be found in the Deutscher Arzneimittel Codex (DAC), Codex Probe no. 5. The particulate matter is evaluated for particles which can be seen easily, well, or with difficulties.

For instance:

- No visible particle: no point
- Particle difficult to be seen within 5 seconds: one point
- Particle easily to be seen within 5 seconds: two points
- Particles to be seen immediately and in higher numbers: ten points

The formula for evaluation is: $Q_{TR} = \frac{A}{N}$, where A stands for the number of points recorded by three test persons and N stands for the number of controlled containers.

The results of all controls for one batch and from batch to batch is very important for the evaluation of the quality and the release for distribution. Trends for a homogeneous or heterogeneous process should be addressed and recognized as soon as it happens. Statistical evaluation of all available data is essential for the routine evaluation of the drug quality.

10.4 META-ANALYTIC MODELING OF DATA OCCURRING IN QUALITY ASSURANCE

Very often quality assurance is based on the availability of a number of batches each having a certain number of items. For example, we might consider again Q_{TR} and define X as

$$X = \text{Number of times with } Q_{TR} \text{ positive in a series of } n \text{ investigations.}$$

This is best demonstrated by means of an example which is taken from the book of Derman and Ross (1997). The data are provided in Table 10.1 and visualized by means of a confidence interval plot (proportion with 95% confidence interval) in Figure 10.1.

Table 10.1 Number of Defective Items for 20 Batches of 200 Items Each

Batch	Number of Defectives	Batch	Number of Defectives
1	24	11	4
2	22	12	13
3	12	13	17
4	13	14	5
5	15	15	9
6	11	16	0
7	25	17	19
8	16	18	0
9	23	19	22
10	14	20	17

As has been pointed out in the literature (Petitti, 1994), the area of meta-analysis has received various impulses during its historic development. In psychology, the development of measures was achieved which could be suitably used for meta-analysis such as the standardized effect difference. Another impulse was the development of suitable statistical methods such as the appropriate form of a *pooled mean*. Meta-analysis experienced tremendous impulses by means of embedding important application areas such as evaluation research or health reporting. It is hoped that both areas discussed in this chapter, namely quality control and assurance and meta-analysis, experience a similar impulse from each other.

It is quite obvious that in quality control the single batch can play the role of a single study in conventional meta-analysis. This can avoid various techniques including control charts and repeated testing, which can be statistically flawed. For example, if 20 binomial tests are employed for the data provided in Table 10.1, it can be expected that one of these will show a significant deviation from a desired standard though there is in fact no deviation from the desired standard (process is still in control). Similarly, if control charts are used, it is well-known that the boundaries of these charts are reached for some batch, though the process is still in control. As a consequence, investigators in quality assurance are forced to investigate for a non-existing source of deviation of the production process.

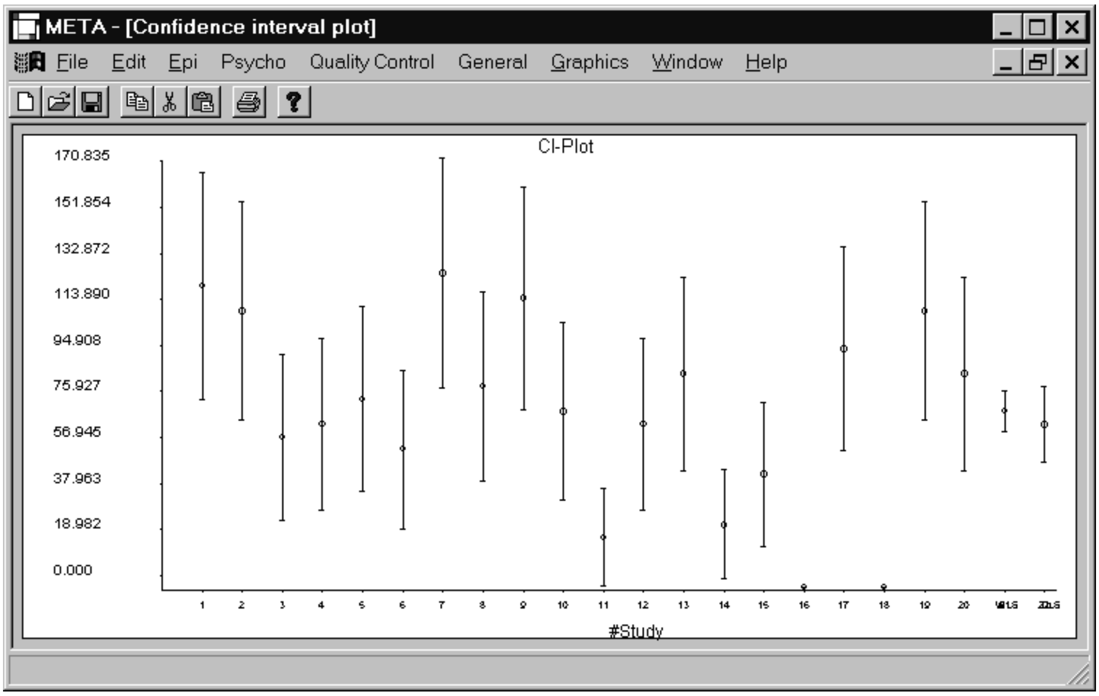


Figure 10.1 Confidence interval plot from the package META for a textbook example of proportion of defective items for 20 batches with 200 items each.

10.5 THE PROBLEM OF HETEROGENEITY

In fact, we are interested in separating *random deviations*, which are occurring always in non-deterministic systems[1], and *systematic deviations*. Only the latter are relevant and prone for further investigation and research.

How can we accomplish this separation? The first step is to model the situation when the process is in control, which is called the situation of *homogeneity*. Typically, it is possible to derive some probability distribution for the measure of interest under homogeneity. We call the associated density of the measure of interest X: $f(x,\theta)$, where θ is some parameter involved in this density. In our example, the *number of defective items*, X, follows a binomial distribution with density $f(x,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, where n is the size of the batch and the parameter θ corresponds to the allowed number of defectives.

The question at hand is: What will happen if a deviation (loss in quality) occurs and how is this reflected in the statistical model? Clearly, if this happens, homogeneity conditions no longer hold and the simple statistical model $f(x,\theta)$ will no longer be correct.

There are some simple tests available which allow to diagnose this situation rather quickly. One of these tests is based upon the defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - E(X_i))^2}{\mathrm{Var}(X_i)}.$$

Typically, $E(X_i)$ and $\mathrm{Var}(X_i)$ will be functions of the unknown parameter θ and plug-in-estimates must be utilized. These plug-in estimators must be constructed with care to achieve χ^2-distribution under homogeneity, at least approximately. To give a demonstration, we note that in our binomial quality control example $E(X_i) = n\theta_i$ and $\mathrm{Var}(X_i) = n\theta_i(1-\theta_i)$, which might lead to the plug-in estimates $\widehat{E(X_i)} = X_i$ and $\widehat{\mathrm{Var}(X_i)} = X_i(1 - X_i/n)$. It can be shown that the associated distribution under homogeneity is quite different from a χ^2-distribution with $k-1$ degrees of freedom if sample sizes per batch, n, are small or moderate, even if the number of batches k becomes large. The right thing to do here turns out to be a variance estimate utilizing information from all batches: $\widehat{\mathrm{Var}(X_i)} = S_k(1 - S_k/n)$, where $S_k = \sum_{i=1}^{k} X_i/k$. The associated χ^2-statistic (with $E(X_i) = S_k$) can be shown to be validly approximated by a χ^2-distribution with $k-1$ degrees of freedom even for small batch size n (like $n = 5$). For further discussion, see Böhning (2000b) as well as Hartung and Knapp (Chapter 4, this volume). To finish this aspect, we find a value of $\chi^2 = 70.41$ with 19 degrees of freedom for the data of Table 10.1, which indicates strongly the presence of heterogeneity.

[1] The question which system is deterministic and which is not is a mere philosophical question. Our point of view is that it is appropriate and useful to consider stochastic variation even when measurements and processes are done with the highest accuracy.

In the following section we will concentrate on the aspect: What can be done if heterogeneity is present?

10.6 MODELING HETEROGENEITY USING MIXTURE DISTRIBUTIONS

If heterogeneity is present it is implied that the proportion of defectives in the batch is deviating in a systematic way from the required standard, in other words, it can be assumed that the hypothesis $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$ is wrong and it is more reasonable to assume that for certain parts of the population of all possible batches a value (for the proportion of defectives) of θ_1 – for other parts a value of θ_2 – is valid and so forth. That is, the population of possible batches consists of a proportion p_j of batches with θ_j, for $j = 1, \ldots, k$. It can be shown (Böhning, 2000a) that in this situation X_i has a mixture distribution

$$f(x_i, P) = \sum_{j=1}^{k} f(x_i, \theta_j) p_j$$

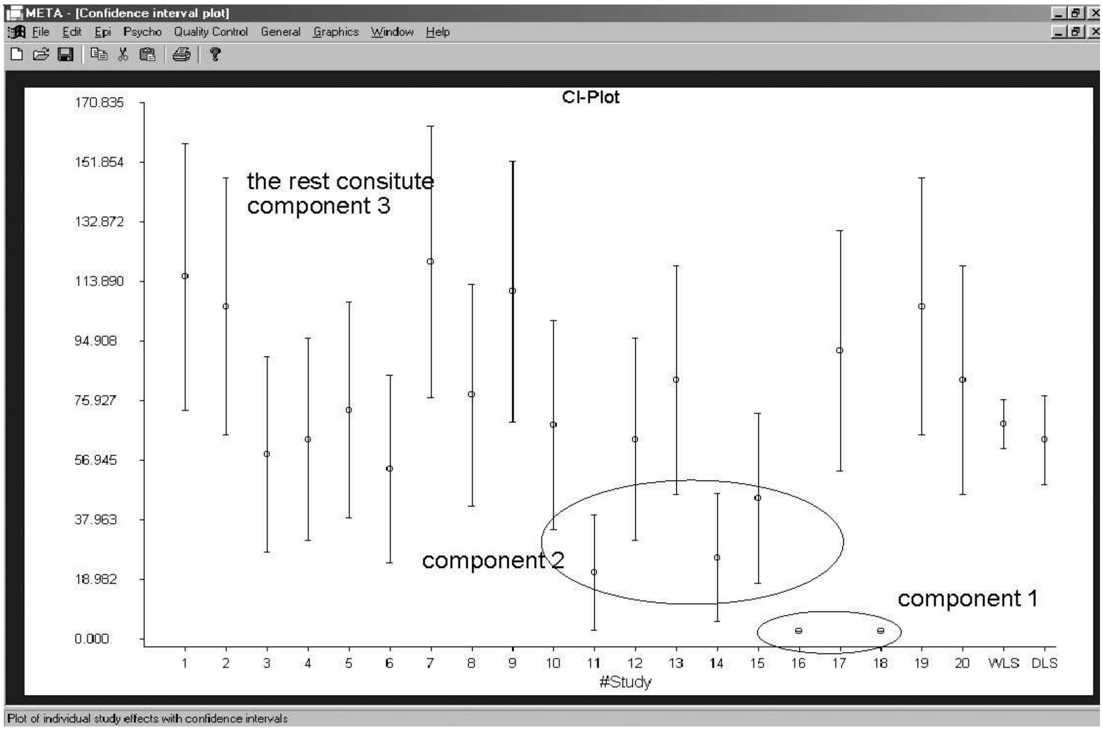
which takes the form of a mixture of binomial distributions for our textbook example:

$$f(x_i, P) = \sum_{j=1}^{k} \binom{n}{x_i} \theta_j^{x_i} (1 - \theta_j)^{n - x_i} p_j. \tag{10.1}$$

The distribution which gives probability mass p_j to θ_j is called *mixing distribution* and is denoted by P. To estimate the parameters involved in Equation 10.1, in other words the mixing distribution P, we use maximum likelihood estimation including the number of components in the mixture k. This can be accomplished with the computer package `C.A.MAN` (see Böhning, Schlattmann, & Lindsay, 1992; Böhning, Dietz, & Schlattmann, 1998). The associated maximum likelihood estimate of k and θ_j, p_j for $j = 1, \ldots, k$ is called *nonparametric maximum likelihood estimate* (NPMLE) of the mixing distribution P. It is usually advisable to check whether the number of components k can be reduced, which can be accomplished by comparing log-likelihoods for reduced values of k such as $k - 1, k - 2, \ldots$ until no significant drop in the log-likelihood is notable. For these fixed values of k estimation is done via the EM-algorithm (Dempster, Laird, & Rubin, 1977).

To provide a demonstration for this technique, we study the data of Table 10.1 again and use the mixture model provided in Equation 10.1. Table 10.2 provides the results. There is empirical evidence for *heterogeneity* and that this heterogeneity consists of 3 components.

It can be seen that the population of batches can be separated into *three* components. One component consists of batches which are free of defective items (9.9%). The second component has 2.87 defective items per 100 (13.3%), whereas the last one has 8.6 defective items per 100, representing the majority of all batches (76.8%).


META - [Confidence interval plot]
File Edit Epi Psycho Quality Control General Graphics Window Help
CI-Plot
170.835
151.854
132.872
113.890
94.908
75.927
56.945
37.963
18.982
0.000
the rest consitute component 3
component 2
component 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 WLS DLS
#Study

Plot of individual study effects with confidence intervals

Figure 10.2 Classification of the batches into their associated components for the textbook example of proportion of defective items for 20 batches with 200 items each.

Table 10.2 Identification of Heterogeneity Structure for 20 Batches of 200 Items Each

Number of Components k	Log-Likelihood
4 (NPMLE)	−63.1454
3	−64.0984
2	−70.9835

Estimated Mixing Distribution for $k = 3$

Proportion θ_j	Weight p_j
0.0000	0.0996
0.0287	0.1326
0.0865	0.7678

Finally, it is even possible to allocate each observed (investigated) batch to one of the components in the mixture. This can be accomplished by utilizing Bayes theorem and calculate the posterior distribution of θ as

$$f(\theta_j|x_i) = \frac{f(x_i, \hat{\theta}_j)\hat{p}_j}{\sum_{l=1}^{k} f(x_i, \hat{\theta}_l)\hat{p}_l},$$

where $\hat{\theta}_j$ and $\hat{p}_j$ correspond to the maximum likelihood estimates identified in the previous estimation process. Each batch i with number of defectives X_i is allocated to that component j for which $f(\theta_j|x_i)$ is largest of all $j = 1, \ldots, k$. This is done for the data in Table 10.1 and the results are provided in Table 10.3. Figure 10.2 also visualizes this reclassification. This technique might enable the practitioner to search for common sources for the occurred heterogeneity and finally identify sources for the loss in quality standards.

Table 10.3 Classification of Each Batch Into the Components

Batch i	X_i	Component j	Batch i	X_i	Component j
1	24	3	11	4	2
2	22	3	12	13	3
3	12	3	13	17	3
4	13	3	14	5	2
5	15	3	15	9	2
6	11	3	16	0	1
7	25	3	17	19	3
8	16	3	18	0	1
9	23	3	19	22	3
10	14	3	20	17	3

10.7 DISCUSSION

We touched upon an approach which explicitly allows the modeling of heterogeneity. To do this, it is important to emphasize that an appropriate measure of interest (describing the quality standards) has to be chosen. Given the chosen measure of interest, it is furthermore equally important to find the corresponding statistical model under homogeneity conditions and further the associated mixture model which models potential heterogeneity. A variety of situations have been assembled to form a package `META` which allows the user in a simple way to analyze heterogeneity problems in his/her application. For details, see Schlattmann, Malzahn, and Böhning (Chapter 16, this volume).

REFERENCES

Böhning, D. (2000a). *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping, and others.* Boca Raton, FL: Chapman & Hall/CRC.

Böhning, D. (2000b). *The flaw in χ^2-heterogeneity tests: A revisit of the Neyman-Scott problem?* Unpublished report, Free University Berlin.

Böhning, D., Dietz, E., & Schlattmann, P. (1998). Recent developments in computer-assisted analysis of mixtures. *Biometrics, 54*, 525–536.

Böhning, D., Schlattmann, P., & Lindsay, B. G. (1992). Computer-assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics*, *48*, 283–303.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Derman, C., & Ross, S. M. (1997). *Statistical aspects of quality control.* San Diego, CA: Academic Press.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Petitti, D. B. (1994). *Meta-analysis, decision analysis and cost-effectiveness analysis. Methods for quantitative synthesis in medicine.* Oxford: Oxford University Press.

Acknowledgments

This research was done under support of the *Bundesministerium für Bildung und Wissenschaft, Forschung und Technologie (BMBF).*

11
Influential Factors for Sensitivity and Specificity for Serodiagnosis of Human and Porcine Trichinellosis

Matthias Greiner
International EpiLab
Danish Veterinary Institute, Copenhagen

Karl Wegscheider
Institute for Statistics and Econometrics, Department of Economy
University of Hamburg

Dankmar Böhning
Working Group: Biometry and Epidemiology
Institute for International Health, Joint Center for Humanities and Health Sciences
Humboldt-University in Berlin / Free University Berlin

Susanne Dahms
Institute for Biometrics and Information Processing
Free University Berlin

Summary

The diagnostic sensitivity and specificity of enzyme-linked immunosorbent assays (ELISAs) for the detection of *Trichinella* antibodies in humans and swine was assessed by a systematic, quantitative literature review. The objective was to identify influential factors for specificity and sensitivity covering a wide range of technical and study design characteristics. Nine out of 12 publications selected for analysis reported more than one pair of sensitivity/specificity. We suggest an explorative-analytical approach that accounts for these "multiple-study type" publications. Two mixed logistic regression models that included study specific explanatory variables, an adjustment for the cut-off value (both fixed effects) and a random effects term (publication) were established for analysis of specificity and sensitivity. The use of an elaborated test antigen was associated with perfect (100%) specificity. In studies that used crude antigen preparations, a positive effect on specificity was associated with publication after 1991, application of the test for humans (versus swine), single-point (versus titration) assays and testing of healthy or non-target (versus other) populations. A positive effect on sensitivity was associated with application of the test for swine (versus humans), testing of other populations than experimentally infected swine or advanced human cases, testing after 26 (versus less than 26) days post infection and medium (versus small, $n < 16$) sample sizes. The impact of the sample size and the status of the positive reference population is obscure and may be due to uncontrolled confounding. The other effects are plausible and show that this form of "exploratory meta-analysis" of diagnostic tests is of practical concern.

11.1 INTRODUCTION

The evidence for the accuracy of a diagnostic test is usually based on multiple primary validation studies rather than on a single study. Multiple studies cover a wider range of marginal conditions such as reference populations, study design and laboratory proficiency and, therefore, are thought of yielding more reliable test performance parameters. The planned multi-centre validation study and the systematic review of published studies are important realizations of a multiple-study based test validation and differ in the extent to which the involved primary studies can be controlled for marginal conditions. Various methods are described for a quantitative summary of multiple validation studies which is here referred to as meta-analysis of diagnostic tests (MADT). These methods include the summary receiver operating characteristic (sROC) analysis (Hurblut III, Littenberg, & Diagnostic Technology Assessment Consortium, 1991; Moses, Shapiro, & Littenberg, 1993), weighted mean values of sensitivity and specificity (Carlson, Skates, & Singer, 1994), relative risk (Mantha et al., 1994), and standardized mean difference (Hasselblad & Hedges, 1995). Irwig et al. (1994) pointed out that a simple pooling of sensitivity (*Se*) and specificity (*Sp*) estimates across primary studies is not appropriate because this would underestimate the overall accuracy. The methods for MADT usually presuppose that each primary study contributes exactly one pair of *Se* and *Sp*, that is, one data point in the ROC space. We refer this to as *single study type*. The data points are further assumed to be independent. In practice we are concerned with deviations from this ideal situation since published evaluation studies often provide more than one estimate of *Se* and *Sp*. We refer such studies to as *multiple-study type* and distinguish three cases. Firstly, more than one test entity (i.e., different tests or technical modifications or application of one test to different host species) is described in a single publication (*multiple-study type I*). Secondly, a set of different cut-off values is used for evaluation (*multiple-study type II*). Thirdly, multiple reference populations are used (*multiple-study type III*). We further distinguish between an enrollment of distinct (mutually exclusive and independent) reference populations (*multiple-study type IIIa*) and repeated measurements on the same reference population (*multiple-study type IIIb*). Combinations of various multiple-study types may occur. We do not consider the case of multiple reference methods and argue for the selection of the most reliable (in terms of accuracy) method as gold standard instead. The scope of a MADT is usually restricted to a single test entity but situations may occur in which a comparison of the test performance between different test entities is relevant (Irwig, Macaskill, Glasziou, & Fahey, 1995). In analogy with the general meta-analytic terminology we shall refer the estimate of the diagnostic test performance to as effect size.

In this chapter, we describe a meta-analytic approach for the validation of diagnostic tests when the source data include multiple-study type publications. Our data derive from a systematic review of published studies on the validation of ELISAs for the detection of *Trichinella* antibodies in humans and swine. Trichinellosis is a zoonosis with severe medical implications if untreated. The ELISA is recommended for diagnosis of both human (Ljungstrom, 1983) and porcine infection (Gamble, 1997). Furthermore, ELISA testing may become mandatory for certification of "*Trichinella* free" pig production within the framework of an anticipated modified trichinellosis control scheme in countries of the European Union (Borowka & Ring, 1997). The emphasis of our application is to identify influential factors for the diagnostic accuracy covering a broad range of marginal conditions rather than validation of a single test entity.

11.2 MATERIALS AND METHODS

11.2.1 Literature Retrieval

The databases Medline™, VetCD™, BeastCD™, and CAB Helminthological Abstracts™ were used as searching frame as described elsewhere (Greiner, Böhning, & Dahms, 1997). The list of retrieved publications was cross-checked and supplemented by experts (Dr. K. Nöckler and Dr. W. P. Voigt, Federal Institute for Health Protection of Consumers and Veterinary Medicine, Berlin, and Dr. K. Wacker, Institute for Epidemiological Diagnosis, Federal Institute for Virus Diseases of Animals, Wusterhausen). Included for analysis were studies on trichinellosis antibody ELISAs in humans and farm pigs published from 1990 to 1995, where the number of true positive, false positive, false negative and true negative test results was either indicated or derivable from the published data. Furthermore, inclusion required a minimum of 5 subjects for each reference sample. Positive subpopulations were not considered if sampled before day 10 post infection or, in the case of repeated measurements (multiple-study type IIIb), earlier than 10 days after the preceding sampling date. Generally, language was not an exclusion criterion except Chinese without translation. A list of publications excluded from this study can be obtained from the first author.

11.2.2 Data Transcription

One data base was constructed that included all available estimates of specificity. We considered here (where applicable) different test entities (i.e., distinct technical procedures) described in one publication as well as different applied cut-off values, different negative subpopulations tested and different time points at which the negative subpopulation was sampled for one test entity, respectively. Basic outcome variables were the number of true negative observations (TN) and the sample size of the respective negative reference population (NNEG), respectively (we omit the index for the unit of observation). We considered the sensitivity associated with one unit of observation as the weighted (using the sample size) average of all available sensitivity estimates with the respective test entity. In case of repeated measurements, we selected the first sampling date following the 35th day after infection as base for sensitivity estimation. A second data base was constructed analogously and comprised all available estimates of sensitivity. Basic outcome variables were the number of true positive observations (TP) and the sample size of the respective positive reference population (NPOS). We considered the specificity associated with one unit of observation as the weighted (using the sample size) average of all available specificity estimates with the respective test entity.

A set of variables was recorded as covariate information for each study. The publication YEAR (0 = 1990, 1991, 1 = 1992+) was recorded from the bibliographic data. The variable SPECIES (0 = human, 1 = swine) denotes the species tested. Variables describing technical aspects were AGPREP (coating antigen; 0 = crude preparation or extract of larval antigen, 1 = excretory/secretory (E/S) antigen or purified preparations), CONJUG (specificity of the anti species-enzyme conjugate; 0 = anti whole-Ig fraction, 1 = anti IgM, IgG, or IgE fraction), TITER (dilution of serum samples; 0 = single-point determination, 1 = titration). The selection of a cut-off value in favor of specificity (e.g., the confidence limit of the negative reference population) was coded with SPW (specificity optimized; 0 = no, 1 = yes). Design characteristics were recorded by the variables STATN (status of the negative reference population; 0 = healthy controls, samples from a non-target population or unrelated diseases, e.g., atopic conditions, 1 = any other selection), STATP (status of the positive reference population; 0 = experimental infection or extreme cases, 1 = any other selection), DPI (days post infection at which the positive reference population was sampled; 0 = 10–25, 1 = 26+, 2 = no information), NNEG and NPOS (categorized sample size for the positive and negative reference population, respectively; 0 = 5–15, 1 = 16–50, 2 = 51+) and RESUBST (identity of the reference population for cut-off selection and test validation; 0 = no, 1 = yes).

11.2.3 Analysis of Influential Factors for Specificity and Sensitivity

The analysis of influential factors on specificity and sensitivity was addressed by two mixed logistic-binomial regression models for distinguishable data that take into account the correlation within-publications. The fixed effects term includes the intercept (α_F), the logit transformation of the associated "counter-parameter" (x_i) and the row-vector of explanatory factors ($\mathbf{z}_i$). The random effects term consists of the random effects parameter (α_R). The data were matched on the publication. The models have the general form

$$\text{logit}(p_i) = \alpha_F + \alpha_R + \beta x_i + \gamma^T \mathbf{z}_i.$$

For analysis of specificity, p_i denotes the simple proportion of $\text{TN}_i/\text{NNEG}_i$, X_i denotes the logit transformation of the associated sensitivity (with $^1/_2$ correction) and $\mathbf{z}_i$ denotes a row-vector of explanatory factors. The variables YEAR, SPECIES, CONJUG, TITER, SPW, STATN, NNEG and RESUBST were selected as candidates for explanatory variables in the analysis of specificity. Multilevel variables were used after dummy coding. The inclusion of the counter-parameter takes care for the part of the variance that can be explained by the applied cut-off value. $\hat{\alpha}_F$, $\hat{\alpha}_R$, $\hat{\beta}$ and the vector $\hat{\gamma}^T$ are empirical coefficients and were found with standard algorithms (logistic-binomial model for distinguishable data with 6 support points; EGRET LBDD(6) module; Statistics and Epidemiology Research Corporation (SERC), 1988). A stepwise backwards fitting strategy was used whereby the variable with the highest p-value of the likelihood ratio statistic (LRS = deviance without/with variable, referred to the chi-square distribution with degrees of freedom, df = number of levels of the variables minus 1) was excluded. This procedure was repeated until the LRS was significant ($p < .05$) for all variables. The goodness-of-fit of the final model was assessed by the LRS of the final model. The candidate variables were also analyzed in univariate mixed logistic regression models. The analysis of sensitivity was accomplished in a complementary manner. Candidates for explanatory variables were YEAR, SPECIES, AGPREP, CONJUG, TITER, STATP, DPI, NPOS, and RESUBST.

11.2.4 Further Analyses

The effect sizes in terms of sensitivity and specificity of the trichinellosis ELISAs were displayed in the ROC space to visualize the scatter of the estimates (Figure 11.1). All possible combinations of a sensitivity and a specificity estimate were considered in case of multiple-type studies. A summary ROC function was established as described by Moses et al. (1993). A chi-square test on homogeneity ($\alpha = .05$; df= number of estimates minus 1) of sensitivity and specificity estimates was done using TP and TN as observed frequencies and NPOS$\times\widehat{Se}_p$ and NNEG$\times\widehat{Sp}_p$ as expected frequencies, respectively (Stata macro "chitest" by Nick Cox, personal communication, StataCorp, 1997). Here, $\widehat{Se}_p$ and $\widehat{Sp}_p$ denote the pooled sensitivity and specificity, respectively. Using a total of r estimates of sensitivity,

$$\widehat{Se}_p = \frac{\sum\limits_{i=1}^{r}(\mathrm{TP}_i)}{\sum\limits_{i=1}^{r}(\mathrm{NPOS}_i)}$$

and using s estimates of specificity,

$$\widehat{Sp}_p = \frac{\sum\limits_{i=1}^{s}(\mathrm{TN}_i)}{\sum\limits_{i=1}^{s}(\mathrm{NNEG}_i)},$$

respectively.

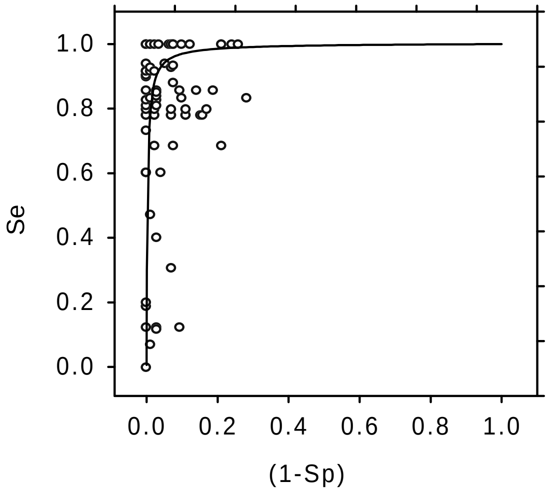


Figure 11.1 Summary ROC plot of the diagnostic specificity (Sp) and sensitivity (Se) of *Trichinella* antibody ELISAs (meta-analysis of 12 studies published between 1990 and 1995). The points represent the reported pairs ($\widehat{Se}, \widehat{Sp}$) in case of a single-type publication and all possible combinations of the two estimates reported for one test entity in case of multiple-type publication (refer to the text for further explanation). The summary ROC function is displayed as solid line.

11.3 RESULTS

11.3.1 Data Transcription

The data from twelve publications (7 on human and 5 on porcine trichinellosis) were included in this meta-analysis (Arriaga, Yepez–Mulia, Morilla, & Ortega–Pierres, 1995; Bruschi, Tassi, & Pozio, 1990; Chan & Ko, 1990; Dzebenski, Bitkowska, & Plonka, 1994; Gamble, 1995; Lind et al., 1991; Mahannop, Chaicumpa, Setasuban, Morakote, & Tapchaisri, 1992; Mahannop, Setasuban, Morakote, Tapchaisri, & Chaicumpa, 1995; Morakote et al., 1991; Morakote, Sukhavat, Siriprasert, Suphawitayanukul, & Thamasonthi, 1992; Nöckler, Voigt, Protz, Miko, & Ziedler, 1995; Serrano, Perez, Reina, & Navarrete, 1992). Three studies belonged to the single study type, nine studies belonged to one of the multiple-study types (Table 11.1). The null hypothesis of homogeneity of the specificity estimates could not be rejected ($\chi^2 = 3.87$; $df = 33$, $p = 1.0$). The null hypothesis of homogeneity of the sensitivity estimates was rejected ($\chi^2 = 132.53$; $df = 55$, $p < .001$). The distribution of study characteristics (here referred to as covariate factors) is described elsewhere (Greiner et al., 1997).

Table 11.1 Types of Evaluation Studies of *Trichinella* Antibody ELISAs Published Between 1990 and 1995 and Number of Analytical Units They Contribute to the Analysis of Specificity and Sensitivity[a]

PUBNR[b]	Study Type	*m*	*c*	*n*	*p*	*t*	Specificity	Sensitivity
1	I/IIIa	3	1	4	1	1	12	3
2	IIIa	1	1	2	1	1	2	1
3	IIIa	1	1	2	1	1	2	1
4	I/IIIb	3	1	1	1	7	3	21
5	I/IIIa/IIIb	2	1	2	1	2	4	4
6	single	1	1	1	1	1	1	1
7	I	3	1	1	1	1	3	3
8	IIIa	1	1	2	1	1	2	1
9	single	1	1	1	1	1	1	1
10	II/IIIb	1	2	1	1	5	2	10
11	IIIb	1	1	1	1	9	1	9
12	single	1	1	1	1	1	1	1
Total							34	56

Note. [a] A published primary study that reports only one estimate of sensitivity and specificity is referred to as single study type. A publication for which one or more of the values m, c, n, p and t is greater than 1 is referred to as a multiple study, where m is the number of test entities ($m > 1$ for multiple-study type I), c is the number of cut-off values ($c > 1$ for multiple-study type II), n and p is the number of negative and positive reference populations considered, respectively ($n + p > 2$ for multiple-study type IIIa), and t is the number of time points at which the positive reference populations was evaluated ($t > 1$ for multiple-study type IIIb).

[b] PUBNR=publication number.

11.3.2 Influential Factors for Sensitivity and Specificity

We found a wide range of estimates for specificity and particularly for sensitivity. The variability was not only due to different cut-off values as suggested by the deviations of data points from the summary ROC function (Figure 11.1). The interest was to identify the main reasons for this variability. An overwhelming positive effect of AGPREP on specificity was observed (Table 11.2). Therefore, sub-studies that used an elaborated antigen (AGPREP = 1) were excluded from the analysis of further explanatory factors for specificity.

Table 11.2 Impact of the Type of Antigen Preparation (Crude and Elaborated) Used in *Trichinella* Antibody ELISAs on the Test Specificity ($\widehat{Sp}$) Based on 12 Studies Published Between 1990 and 1995

	Crude (AGPREP = 0)	Elaborated (AGPREP = 1)
$\widehat{Sp} < 1$	19	0
$\widehat{Sp} = 1$	8	7

Using a stepwise backward fitting procedure of a mixed logistic regression model, we identified four variables as potential factors for specificity. Four other candidate variables (CONJ, SPW, NNEG, RESUBST) were excluded due to non-significant LRSs. The extension of the base model that included the fixed effects intercept, the counter parameter, and the random effects term by the four explanatory variables was significant (LRS ($df = 4$) = 31.4, $p < .001$). According to the (Wald test significant) effects in the final model, the specificity appeared to be better in studies published after 1991 (YEAR, $p < .001$), better in humans than in swine (SPECIES, $p = .010$), better in single-point assays than in titration assays (TITER, $p < .001$), and better in healthy, non-target or unrelated reference controls than in any other control samples (STATN, $p = .001$). The counter-parameter (logit Se) had a significant ($p < .001$) negative effect (Table 11.3).

The inferences from univariate analyses were consistent for two (TITER, STATN) variables. The effects of YEAR and SPECIES were not discovered whereas NNEG = 1 and RESUBST were associated with a positive and negative univariate effect. For the analysis of potential factors for sensitivity, we selected (stepwise backward fitting procedure) four variables as potential factors for specificity. Five other candidates for sensitivity analysis (YEAR, ANTIG, CONJ, TITER, RESUBST) were excluded due to non-significant LRSs. The extension of the base model by the four explanatory variables was significant (LRS($df = 6$) = 321.6, $p < .001$). According to the (Wald test significant) effects in the final model, the sensitivity appeared to be better in swine than in humans (SPECIES, $p = .005$), better in any other than extreme cases or experimental infections (STATP, $p = .005$), better when samples where taken 26 days or more after infection (DPI = 1, $p < .001$), and better in sub-studies that used a sample size between 16 and 50 in than smaller sub-studies (NPOS = 1, $p = .006$). The counter-parameter (logit Sp) had a significant ($p < .001$) neg-

Table 11.3 Coefficients (and Wald Test p-values) From Univariate and Multivariate Mixed Effects Logistic Regression Models for Analysis of Explanatory Variables for the Diagnostic Specificity and Sensitivity of *Trichinella* Antibody ELISAs (Meta-Analysis of 12 Studies Published Between 1990 and 1995)[a]

Variable[b]	Specificity (Sp)				Sensitivity (Se)			
	Univariate		Multivariate		Univariate		Multivariate	
YEAR	0.31	(.439)	1.17	(< .001)	0.08	(.610)	n.i.	
SPECIES	0.71	(.258)	−1.67	(.010)	0.22	(.119)	4.25	(.005)
AGPREP	n.i.		n.i.		0.86	(.001)	n.i.	
CONJUG	−17.40	(.999)	n.i.		−1.58	(.061)	n.i.	
TITER	−1.08	(.018)	−1.70	(< .001)	0.02	(.936)	n.i.	
SPW	0.21	(.637)	n.i.		n.a.		n.a.	
STATN	−2.07	(.006)	−2.43	(.001)	n.a.		n.a.	
STATP	n.a.		n.a.		−0.19	(.162)	4.25	(.005)
DPI	n.a.		n.a.		21.3	(.995)	20.38	(.995)
					3.5	(<.001)	3.65	(< .001)
NPOS	n.a.		n.a.		−0.15	(.620)	0.23	(.455)
					0	(.994)	0.96	(.006)
RESUBST	−0.56	(.046)	n.i.		0.32	(.461)	n.i.	
X	n.a.		−0.57	(< .001)	n.a.		−0.27	(< .001)

Note. n.i. = variable not included; n.a. = variable not applicable.

[a] Final multivariate models (Sp: $n = 27$; Se: $n = 56$) obtained by stepwise backwards fitting starting. Base models included intercept term, counter parameter (logit($\widehat{Se}$) and logit ($\widehat{Sp}$) for analysis of Sp and Se analysis, respectively), and a random effects (RE) term. The coefficient for the RE term was 0.23×10^{-14}, and 0.95 for the Sp and Se model, respectively.

[b] YEAR, publication year (0 = 1990, 1991, 1 = 1992+); SPECIES (0 = human, 1 = swine); AGPREP, coating antigen (0 = crude preparation or extract of larval antigen, 1 = excretory/secretory antigen or purified preparations); CONJUG, specificity of the anti species-enzyme conjugate (0 = anti whole-Ig fraction, 1 = anti IgM, IgG, or IgE); TITER, dilution of serum samples (0 = no titration, 1 = titration); SPW, specificity optimized (0 = no, 1 = yes); STATN, status of the negative reference population (RP) (0 = healthy controls, samples from a non-target population or unrelated diseases, 1 = any other selection); STATP, status of the positive RP (0 = experimental infection or extreme cases, 1 = any other selection); DPI, days post infection (0 = 10–25, 1 = 26+, 2 = no information); NNEG and NPOS, categorized sample size for the positive and negative RP, respectively (0 = 5–15, 1 = 16–50, 2 = 51+); RESUBST, identity of the RP for cut-off selection and validation (0 = no, 1 = yes); X = counter parameter. The base category is 0 for all variables. For variables with more than two categories, the coefficients are shown for categories in decreasing order.

ative effect (Table 11.3). By univariate analysis we found a consistent effect of DPI = 1 whereas discrepant results were obtained for SPECIES, STATP and NPOS = 1 (no effects) and for AGPREP (positive effect).

11.4 DISCUSSION

11.4.1 Parameter Heterogeneity and the Impact of Influential Covariate Factors

The objective of the study was to analyze influential factors for specificity and sensitivity of trichinellosis antibody ELISAs based on a systematic review of the literature. Our intention was not to estimate the "global diagnostic accuracy" of trichinellosis serology. Such an enterprise was strictly invalid due to the inclusion of different test systems in our review. The wide range of marginal conditions in the published studies a priori justified the assumption of parameter heterogeneity (i.e., differences in the diagnostic accuracy between studies). In fact, homogeneity could be rejected for sensitivity but not for specificity. However, since the power of homogeneity tests is generally limited, we continued to investigate explanatory factors for sensitivity and specificity using two separate logistic regression models.

11.4.2 Problem of Multiple Sub-Studies per Publication

Typically, test validation studies are pre-stratified in their design. That means, sample sizes and population characteristics are pre-determined and, consequently, Se and Sp are stochastically independent random variables. A problem arises when multiple estimates of Se and/or Sp are reported in primary studies, for example, when more than two reference samples (multiple-study type IIIa) or repeated measurements (multiple-study type IIIb) are encountered. Obviously, the pooling of type III study data results in a loss of information that may be useful to study influential factors for test accuracy. On the other hand, since meta-analysis generally deals with summary measures of sensitivity and specificity, a complete analysis should involve *all possible combinations* of the reported Se and Sp estimates, which results in a data augmentation. A preliminary analysis of this data set treated these combinations as if they were independent (Greiner et al., 1997). This approach was associated, however, with (i) a substantial violation of the independence assumption, (ii) a severe bias towards significant effects (due to artificially increased sample sizes) and (iii) the risk of bias (with unpredictable direction) due to inappropriate weights. More stringent inclusion criteria and complete avoidance of repeated measurements at the same population were a solution to the problem if the overall goal was to estimate the summary effect size. The separate analysis of specificity and sensitivity may provide a solution to the problem. We have included the counter-parameter into the explanatory part of the models in order to account for the inherent impact of the cut-off value. The lack of independence within publications was considered when we chose a mixed effects model with a random effects term and the publication as matching variable. This approach allows an estimate of effect sizes in the presence of overdispersion (as caused by correlation within publications).

11.4.3 Interpretation of the Multivariate Analyses

We started by postulating that certain covariate factors may be influential for either sensitivity or specificity. Multivariate models that use summary measures (e.g., Moses et al., 1993; Hasselblad & Hedges, 1995) are not suitable to discover such factors. Our analysis overcomes this problem but is limited through the number of published, eligible studies. Petitti (1994, p. 126) argued that a small number of studies should not preclude the application of regression methods, but the number of explanatory variables should be kept small. Using stepwise backwards fitting, eight and nine variables could be investigated simultaneously for analysis of specificity and sensitivity, respectively. The final mixed effects logistic regression models included four explanatory variables (in each case) besides the counter-parameter.

The test specificity was better in tests that used an elaborated antigen as shown by cross-tabulation (Table 11.2). The results of the multivariate analysis of specificity pertain to studies that used a crude or extract antigen preparation (AGPREP = 0). In these studies, the publication year was positively associated with an increase in specificity. Unobserved changes in technical or other factors (as expressed by the surrogate variable YEAR) may have led to a better specificity. Interestingly, a better specificity and worse sensitivity in humans than in swine was found. This finding might reflect a different medical decision making situation. Trichinellosis serology in medicine usually is a confirmatory instrument (with emphasis on specificity) whereas screening applications (with emphasis on sensitivity) are dominating in veterinary applications. This effect cannot be explained by the choice of the cut-off value because the analysis was adjusted for the counter-parameter. The data also suggest that titration methods do not improve the test properties. In fact, according to our results, titration was associated with worse specificity. Single-point ELISAs - preferred for practical and economic reasons - have been recommended for veterinary seroepidemiologic applications (Wright, Nilsson, van Rooij, Lelenta, & Jeggo, 1993). The selection of reference populations is a critical factor in the evaluation process as pointed out elsewhere (e.g., Knottnerus & Leffers, 1992). It is also well recognized that likelihood ratios of diagnostic tests (i.e., combined expressions of Se and Sp used to establish post-test probabilities) are not invariant to changes in the source population (e.g., Miettinen & Caro, 1994). Our results confirm that the specificity may be overestimated when using healthy or non-target populations or patients with unrelated diseases as negative reference population. Experimental infections (in swine) and clinically advanced cases (in humans) were unexpectedly associated with worse sensitivity than other selections of positive reference populations. The opposite seems to be a common finding in laboratory sciences according to Gerhardt and Keller (1986). The duration of infection prior to sampling is related to the degree to which specific antibodies have been produced and, thus, can be considered a true factor for sensitivity. The positive effect of medium sample sizes on the sensitivity is obscure and may be due to uncontrolled confounding. Finally, the negative weights of the counter-parameters included in the models underline the inherent effect of the cut-off value. We had expected that other variables such as the type of immunoglobulin detected with the test would contribute to the explanation of the observed variability of sensitivity and specificity as well. We cannot rule out any of those factors since the number of studies included in our analysis was fairly small. Some of the above mentioned effects were also detected by univariate analysis. However, eight discrepant results show that the lack of adjustment for confounding and interaction may lead to invalid inferences.

11.4.4 Limitations

Some potentially important design factors were not included in the analysis because of their distribution. Blinding, for example, has been suggested as a standard for validation studies (Mulrow, Linn, Gaul, & Pugh, 1989). The knowledge of the true disease status might result in biased (too optimistic) accuracy parameters (“test review bias”; Begg, 1987). Only one (human trichinellosis; PUBNR 1) of the reviewed studies indicated that samples were coded prior to analysis. Furthermore, the diagnostic accuracy will be enhanced if test results within an intermediate range (“grey zone”) were excluded from sensitivity and specificity calculations. Two studies (one on human and one on porcine trichinellosis; PUBNR 6, 9) used intermediate ranges.

11.5 CONCLUSION

The mixed logistic regression models described above have been found suitable to investigate influential factors for specificity and sensitivity of a diagnostic test based on a quantitative, systematic review of the literature. The approach allows the inclusion of studies that contribute more than one pair of parameters.

REFERENCES

Arriaga, C., Yepez–Mulia, L., Morilla, A., & Ortega–Pierres, G. (1995). Detection of circulating trichinella spiralis muscle larva antigens in serum samples of experimentally and naturally infected swine. *Veterinary Parasitology*, *58*, 319–326.

Begg, C. B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine*, *6*, 411–423.

Borowka, H. J., & Ring, C. (1997). Trichinenfreie Region – Eine realistische Prämisse für den Verbraucherschutz [Trichinella-free region – A realistic premise for consumer protection]. *Fleischwirtschaft*, *73*, 1362–1365.

Bruschi, F., Tassi, C., & Pozio, E. (1990). Parasite-specific antibody responses in trichinella sp. human infection: A one year follow-up. *American Journal of Tropical Medicine and Hygiene*, *43*, 186–193.

Carlson, K. J., Skates, S. J., & Singer, D. E. (1994). Screening for ovarian cancer. *Annals of Internal Medicine*, *121*, 124–132.

Chan, S. W., & Ko, R. C. (1990). Serodiagnosis of human trichinosis using a gel filtration antigen and indirect IgG-Elisa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *84*, 721–722.

Dzebenski, T. H., Bitkowska, E., & Plonka, W. (1994). Detection of a circulating parasitic antigen in acute infections with trichinella spiralis: Diagnostic significance of findings. *Zentralblatt Bakteriologie*, *281*, 519–525.

Gamble, H. R. (1995). Detection of trichinellosis in pigs by artificial digestion and enzyme immunoassay. *Journal of Food Protection*, *59*, 295–298.

Gamble, H. R. (1997). Trichinellosis. In Office International des Epizooties (OIE) (Ed.), *Manual of standards for diagnostic tests and vaccines* (pp. 477–480). Paris: OIE.

Gerhardt, W., & Keller, H. (1986). Evaluation of test data from clinical studies. *Scandinavian Journal of Clinical Laboratory Investigations*, *181*, 1–74.

Greiner, M., Böhning, D., & Dahms, S. (1997). Meta-analytic review of ELISA test for the diagnosis of human and porcine trichinellosis: Which factors are involved in diagnostic accuracy? In E. A. Goodall & M. V. Thrusfield (Eds.), *Proceedings of a meeting held at the University College, Chester, 1997* (pp. 12–22). Chester: Society for Veterinary Epidemiology and Preventive Medicine.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167–178.

Hurblut III, T. A., Littenberg, B., & Diagnostic Technology Assessment Consortium. (1991). The diagnostic accuracy of rapid dipstick tests to predict urinary tract infection. *American Journal of Clinical Pathology*, *96*, 582–588.

Irwig, L., Macaskill, P., Glasziou, P., & Fahey, M. (1995). Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology*, *48*, 119–130.

Irwig, L., Tosteson, A. N. A., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T. C., & Mosteller, F. (1994). Guidelines for meta-analysis evaluating diagnostic tests. *Annals of Internal Medicine*, *120*, 667–676.

Knottnerus, J. A., & Leffers, P. (1992). The influence of referral patterns on the characteristics of diagnostic tests. *Journal of Clinical Epidemiology*, *45*, 1143–1154.

Lind, P., Eriksen, L., Henriksen, S. A., Homan, W. L., van Knapen, F., Nansen, P., & Stahl Skov, P. (1991). Diagnostic tests for trichinella spiralis infection in pigs. a comparative study of ELISA for specific antibody and histamine release from blood cells in experimental infections. *Veterinary Parasitology*, *39*, 241–252.

Ljungstrom, I. (1983). Immunodiagnosis in man. In W. C. Campbell (Ed.), *Trichinella and Trichinosis* (pp. 403–424). New York: Plenum Press.

Mahannop, P., Chaicumpa, W., Setasuban, P., Morakote, N., & Tapchaisri, P. (1992). Immunodiagnosis of human trichinosis using excretory-scretory (ES) antigen. *Journal of Helminthology*, *66*, 297–304.

Mahannop, P., Setasuban, P., Morakote, N., Tapchaisri, P., & Chaicumpa, W. (1995). Immunodiagnosis of human trichinosis and identification of specific antigen for Trichinella spiralis. *International Journal of Parasitology*, *25*, 87–94.

Mantha, S., Roizen, M. F., Barnard, J., Thisted, R. A., Ellis, J. E., & Foss, J. (1994). Relative effectiveness of four preoperative tests for predicting adverse cardiac outcomes after vascular surgery: A meta-analysis. *Anesthesia & Analgesia*, *79*, 422–433.

Miettinen, O. S., & Caro, J. J. (1994). Foundations of medical diagnosis – What actually are the parameters involved in Bayes theorem. *Statistics in Medicine*, *13*, 201–209.

Morakote, N., Khamboonruang, C., Siriprasert, V., Suphawitayanukul, S., Marcanantachoti, S., & Thamasonti, W. (1991). The value of enzyme-linked immunosorbent assay (ELISA) for diagnosis of human trichinosis. *Tropical Medicine and Parasitology*, *42*, 172–174.

Morakote, N., Sukhavat, K., Siriprasert, V., Suphawitayanukul, S., & Thamasonthi, W. (1992). Persistence of IgG, IgM, and IgE antibodies in human trichinosis. *Tropical Medicine and Parasitology*, *43*, 167–169.

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnosis test into a summary ROC curve – Data analytic approaches and some additional considerations. *Statistics in Medicine*, *12*, 1293–1316.

Mulrow, C. D., Linn, W. D., Gaul, M. K., & Pugh, J. A. (1989). Assessing quality of a diagnostic test evaluation. *Journal of General Internal Medicine*, *4*, 288–295.

Nöckler, K., Voigt, W. P., Protz, D., Miko, A., & Ziedler, K. (1995). Indirect ELISA for the diagnosis of trichinellosis in living pigs. *Berliner und Münchener Tierärztliche Wochenschrift*, *108*, 167–174.

Petitti, D. B. (1994). *Meta-analysis, decision-analysis, and cost-effectiveness analysis: Methods for quantitative synthesis in medicine.* Oxford: Oxford University Press.

Serrano, F., Perez, E., Reina, D., & Navarrete, I. (1992). Trichinella strain, pig race and other parasitic infections as factors in the reliability of ELISA for the detection of swine trichinellosis. *Parasitology*, *105*, 111–115.

StataCorp. (1997). *Stata statistical software (version 5.0) [Computer software].* College Station, TX: Stata Corporation.

Statistics and Epidemiology Research Corporation (SERC). (1988). *Epidemiological graphics, estimation, and testing package (EGRET). Analysis module (PECAN; version 25.6) [Computer software].* Washington, DC.

Wright, P. F., Nilsson, E., van Rooij, E. M. A., Lelenta, M., & Jeggo, M. H. (1993). Standardisation and validation of enzyme-linked immunosorbent assay techniques for the detection of antibody in infectious disease diagnosis. *Revues Scientifiques et Techniques des Office Internationale des Epizooties*, *12*, 435–450.

Acknowledgments

We thank Dr. K. Nöckler, Dr. W. P. Voigt (Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin, BgVV, Berlin, FRG) and Dr. K. Wacker (Institut für epidemiologische Diagnostik, Wusterhausen, FRG) for cross-checking and supplementation of our literature retrieval. We also thank Dr. N.J. Cox (University of Durham) for making the chi-square test available.

12

Meta-Analysis in Hospital and Clinical Epidemiology based on Mixed Generalized Linear Models

Ekkehart Dietz

Working Group: Biometry and Epidemiology
Institute for International Health, Joint Center for Humanities and Health Sciences
Free University Berlin

Klaus Weist

Institute of Hygiene and Environmental Health
Free University Berlin

Summary

Meta-analyses in the area of hospital and clinical epidemiology have been done for quite some time. Typically, the quantitative part of such analyses is to provide pooled estimators of new hygiene measures or clinical interventions, respectively. If the sizes of the effect in the studies of a meta-analysis were considered to be nearly identical, then the respective pooled estimator could be interpreted as an estimate of the common effect of the new measure. Otherwise, it could be interpreted as an estimate of a mean effect. To draw practical consequences from a mean effect estimation, a description of the heterogeneity of effects is necessary. Such a description is not provided by the standard random effect estimator, which assumes normal distributed study specific effects. As an alternative, a non-parametrical random effect estimator is suggested. This estimator is based on a finite mixed generalized linear model. These models have proven to be very flexible and useful to estimate mean effect sizes and to explain heterogeneity, because they allow for non-normal random

effects and to use covariables to explain baseline and effect heterogeneity for several effect measurements. The method is illustrated using data of two meta-analyses, which have been published recently by Thompson and Sharp (1999) as well as Veenstra, Saint, Saha, Lumley, and Sullivan (1999).

12.1 INTRODUCTION

A main task of hospital epidemiology is to evaluate the hygiene regulations in hospitals with respect to hospital acquired or nosocomial infections. For example, hospital epidemiological studies have to justify certain infection control measures such as modification of central venous catheters to reduce catheter related blood stream infections. For several reasons, hospital epidemiological studies are usually small with respect to the number of patients. Hence, a single study provides only little evidence to indicate that a certain hygiene regulation is better than a standard one. Therefore, meta-analyses in this area have been done for quite some time. The quantitative part of such an analysis is to provide a pooled estimator of the effect of the new hygiene measure. If the sizes of the effect in several studies of a meta-analysis were considered to be nearly identical, such a pooled estimator could be interpreted as an estimate of the *common effect* of the new measure. It could otherwise be interpreted as an estimate of a *mean effect*. To draw practical consequences from a mean effect estimation, a description of the heterogeneity of effects is necessary.

12.1.1 The Data Base

Meta-analyses in hospital epidemiology are typically based on count data obtained in several studies. These are mostly intervention studies. The nature of the count data that can be used does not only depend on the study design but also on the available study report. Two typical types of data layout of studies are shown in Tables 12.1 and 12.2.

Table 12.1 Prevalence Data

Hygiene Regulations	Number of Patients Infected	Number of Patients
Standard	n_0	N_0
New	n_1	N_1
Σ	n	N

Incidence data are published from cohort studies, whereas prevalence data are published either from cohort studies or from cross-sectional studies. In addition to this count data, characteristics of the studies like “year of the study” and “type of hospital” are also available. Mostly, the count data are also available for subgroups of the study populations. Such subgroups are defined by cross classifications by characteristics like “hospital ward”, “gender”, “sever-

Table 12.2 Incidence Data

Hygiene Regulations	Number of Infections	Patient Days
Standard	n_0	N_0
New	n_1	N_1
Σ	n	N

ity of disease", and "age group". These groups are called "units of the study". Their defining characteristics are called "first order variables", whereas characteristics of the studies are called "second order variables". The binary indicator variable "hygiene regulations" (1 = new regulations, 0 = standard regulations) is an example of a first order variable. This is usually the variable of main interest. Of course, it always has to be available. Sometimes, but very rarely, both outcome and explanatory variables of individuals (patients) can be obtained, that is, units are individuals. The meta-analysis can be considered then as the evaluation of a multi-center study.

12.1.2 Effect Measurements and Baseline Heterogeneity

Let

$$R_S = \frac{n_0}{N_0} \quad \text{and} \quad R_N = \frac{n_1}{N_1}$$

denote the infection rates under the standard and the new hygiene regulations, respectively. These two quantities can be used to measure the effectiveness of the new regulations. An overview of measures that are in some use is given in Olkin (1999). The most popular ones are the logarithm of the relative risk ($\log(RR)$), the logarithm of the odds ratio ($\log(OR)$) and the risk difference (RD):

$$\begin{aligned}\log(RR) &= \log(R_N) - \log(R_S)\\ \log(OR) &= \log(R_N/(1-R_N)) - \log(R_S/(1-R_S))\\ RD &= R_S - R_N.\end{aligned}$$

Another often used measure and one which can be derived from the risk difference is the number needed to treat (NNT):

$$NNT = 1/(R_S - R_N).$$

In cohort studies, an intuitively appealing measure of efficacy of new hygiene regulations is

$$ef = RD/R_S. \tag{12.1}$$

It is just the probability that a certain patient who would be infected in a certain period of time under standard hygiene regulations will not be infected in the same period of time under the new hygiene regulations. From Equation 12.1 it follows that

$$\begin{aligned} OR &= \frac{1 - ef - R_S \cdot (1 - ef)}{1 - R_S \cdot (1 - ef)} \\ RR &= 1 - ef, \\ RD &= R_S \cdot ef, \end{aligned}$$

and

$$NNT = 1/(R_S \cdot ef).$$

Thus, one reason for heterogeneity of OR, RD, and NNT in the several studies could be *baseline heterogeneity*, which is the heterogeneity of the rates under standard regulations R_S. Baseline heterogeneity is very common in hospital epidemiologic meta-analyses. The advantage of relative risk RR is that it does not depend on baseline rates. The common efficacy ef can be estimated by the common relative risk RR without considering the baseline heterogeneity and explanatory variables of baseline heterogeneity.

In the case of cross-sectional studies, the odds ratio is preferred. Thereby, the odds ratio is considered as an estimation of the relative risk. This is justified by assuming the mean duration of an infection under the new regulations to be about the same as under the standard regulations. Let D denote the common mean duration of an infection and $R_S^{'}$ and $R_N^{'}$ the underlying incidence under the standard regulations and the new regulations, respectively. Under steady state conditions it holds

$$\frac{R_S}{1 - R_S} = R_S^{'} \cdot D$$

and

$$\frac{R_N}{1 - R_N} = R_N^{'} \cdot D.$$

Therefore,

$$RR^{'} = \frac{R_N^{'}}{R_S^{'}} = \frac{R_N \cdot (1 - R_S)}{(1 - R_N) \cdot R_S} = OR.$$

Thus, in the case of cross-sectional studies, we have an estimation of the odds ratio OR, which is an estimation of the relative risk RR in the study population. To compute asymptotical confidence intervals in the case of small sample sizes, it is advantageous to use $\log(RR)$ and $\log(OR)$ instead of RR and OR, respectively.

12.1.3 Heterogeneity of Effect Size and Standard Methods of Meta-Analysis

Let $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_J$ denote the effect size estimates in the J studies considered and let $w_1, w_2, \ldots, w_J$ denote their respective inverse variances. Common effect size (m) and its standard error (SE) are usually estimated by

$$\hat{m} = \frac{\sum_{j=1}^{J} w_j \hat{b}_j}{\sum_{j=1}^{J} w_j}$$

and

$$SE(\hat{m}) = \left(\sum_{j=1}^{J} w_j \right)^{-\frac{1}{2}},$$

respectively. The null hypothesis $H_0 : m = 0$ is rejected if the absolute value of

$$T = \frac{\sum_{j=1}^{J} w_j \hat{b}_j}{\sqrt{\sum_{j=1}^{J} w_j}}$$

is larger than the $(1 - \alpha)$-quantile of the standard normal distribution, where α is the test level chosen (Thompson, 1993). If the assumption of a common effect size does not hold, a random effect has to be assumed. The standard random effect model is

$$b_j \sim \mathcal{N}(m, \tau^2)$$

and

$$\hat{b}_j \sim \mathcal{N}(b_j, w_j^{-1}).$$

It is assumed that the effect sizes of the selected studies are normally distributed with a certain unknown mean value m and a certain unknown variance τ^2. As a respective estimate of the mean effect size m,

$$\hat{m}^* = \frac{\sum_{j=1}^{J} w_j^* \hat{b}_j}{\sum_{j=1}^{J} w_j^*}$$

is used, where $w_j^* = (w_j^{-1} + \hat{\tau}^2)^{-1}$ and $\hat{\tau}^2$ is a suitable estimator of effect variance τ^2. Confidence intervals and significance tests of the mean effect size can be obtained in a similar fashion as for the common effect estimation. One has simply to replace the w_j by the w_j^* in the respective formulas above. As an estimator of τ^2,

$$\hat{\tau}^2 = \max\left(\frac{Q - J + 1}{\sum w_j - \sum w_j^2 / \sum w_j}\,,\ 0\right) \tag{12.2}$$

can be used (DerSimonian & Laird, 1986). The quantity

$$Q = \sum_{j=1}^{J} w_j(\hat{b}_j - \hat{b})^2$$

in Equation 12.2 can also be used as a test statistic of heterogeneity. If the null hypothesis is true (no heterogeneity), then Q is distributed as χ^2 with $J - 1$ degrees of freedom.

The assumption of a normal distribution of the effect size in the standard random effect model is needed to theoretically justify the mean and variance estimator of the effect size. On the other hand, this assumption provides the "mean effect size" with a certain statistical meaning. However, this assumption is rather restrictive. It is usually not provable, because of the relative small number of studies in a meta-analysis. Consequently, the scientific value and the clinical relevance of the result of meta-analysis based on the standard model are limited. Therefore, more general approaches have been considered recently (Thompson & Sharp, 1999; Aitkin, 1999a; Böhning, 2000a).

In this chapter, a certain generalization of the standard method above is presented and applied to example data. The generalizations are

1 the allowance for non-normal random effects, and

2 the use of covariables to explain baseline and effect heterogeneity.

12.2 THE MODEL

Let y_{ij} denote the value of the count number observed at the ith unit in the jth study, $j = 1, 2, \ldots, J$; $i = 1, 2, \ldots, n_j$. The observations y_{ij} are assumed to be independent random variables having expectations

$$E(y_{ij}) = \mu_{ij}, \quad i = 1, \ldots n_j; \quad j = 1, \ldots, J.$$

In the case of prevalence data

$$y_{ij} \sim \text{Binomial}(\mu_{ij}, N_{ij})$$

is assumed. If the units are individuals, $N_{ij} = 1 \; \forall i, j$. In the case of incidence data

$$y_{ij} \sim \text{Poisson}(\mu_{ij})$$

is assumed. In both cases, the mean parameter μ_{ij} is linked with a linear predictor LP

$$g(\mu_{ij}) = LP_{ij}$$

by a suitable link function $g(\cdot)$, as usual in generalized linear models. For prevalence data (binomial models), we consider the linear predictor

$$LP_{ij} = \beta_1^T X_{ij} + \beta_2^T X_j + z_j + b_j x_{ij},$$

where X_{ij} is a vector of first order variables, X_j is a vector of second order variables, x_{ij} is the binary indicator of the new hygiene regulations (1 = new hygiene regulations, 0 = standard hygiene regulations), and β_1 and β_2 are unknown parameter vectors. z_j and b_j are random effects with a joint distribution

$$\begin{pmatrix} z_j \\ b_j \end{pmatrix} \sim \phi(z, b) \quad \forall j.$$

ϕ remains completely unspecified. The expectation of b_j is the mean effect size m^*, which we are particularly interested in. If there is no effect heterogeneity, then the linear predictor can be simplified by replacing b_j by the fixed effect parameter m in the linear predictor above. If the logit link function $g(\mu_{ij}) = \log(\mu_{ij}/(1 - \mu_{ij}))$ is used, then m^* is just the mean log odds ratio.

This model is very flexible and more general than the standard random effect model of meta-analysis. Baseline heterogeneity is explained by the covariables and by the random effect z_j. Effect heterogeneity is explained by first level covariables and by the random effect b_j of this model. Note, that in case of the logit link the log odds ratio of the r_1th unit versus r_0th unit of the jth study is

$$\text{logit}(\mu_{r_1 j}) - \text{logit}(\mu_{r_0 j}) = \beta_1^T (X_{r_1 j} - X_{r_0 j}) + b_j (x_{r_1 j} - x_{r_0 j}),$$

where the second term of the right side simplifies to b_j if the r_1th unit is a treatment unit and the r_0th unit is a control unit. Another possibility to explain effect heterogeneity is to augment the linear predictor by interaction terms of second level explanatory variables and the treatment indicator x_{ij}.

For Poisson models, we consider the linear predictor

$$LP_{ij} = \beta_1^T X_{ij} + \beta_2^T X_j + z_j + b_j x_{ij} + \log(N_{ij}).$$

The difference to the former linear predictor is the additional offset term $\log(N_{ij})$. Using the log link $g(\cdot) = \log(\cdot)$ leads immediately to

$$m^* = E(\log(RR)).$$

12.3 ML-ESTIMATION

To obtain maximum likelihood estimates of β_1, β_2, and the parameter vector θ of ϕ, the likelihood function

$$L(\beta_1, \beta_2, \theta) = \prod_{j=1}^{J} \int \left\{ \prod_{i=1}^{n_j} f(y_{ij} \mid z_j, b_j, \beta_1, \beta_2, X_j, X_{ij}) \right\} \phi(z_j, b_j) \partial z_j \partial b_j \qquad (12.3)$$

has to be maximized , where

$$f(y_{ij} \mid z_j, b_j, \beta_1, \beta_2, X_j, X_{ij}) = f(y_{ij} \mid LP_{ij})$$

denotes the respective conditional probability density distribution of y_{ij} given the linear predictor. Because we have not specified the distribution $\phi(z_j, b_j)$, we have to look for its nonparametric estimate. For this purpose, it is sufficient to consider two-dimensional discrete distributions with less than $J + 1$ mass points

$$\phi(z, b) = \begin{cases} p_k & \text{if } \begin{pmatrix} z \\ b \end{pmatrix} = \begin{pmatrix} z_k \\ b_k \end{pmatrix}, \\ 0 & \text{otherwise.} \end{cases}$$

$$k = 1, \ldots, K; \quad K \leq J.$$

(see Aitkin, 1999a, 1999b). This distribution has $(3 \cdot K - 1)$ parameters, which are $z = (z_1, z_2, \ldots, z_K)$, $b = (b_1, b_2, \ldots, b_K)$, and $p = (p_1, p_2, \ldots, p_{K-1})$, where $p_K = 1 - \sum_{k=1}^{K-1} p_k$. When using such a distribution, Equation 12.3 simplifies to

$$L(\beta_1, \beta_2, p, b, z) = \prod_{j=1}^{J} \sum_{k=1}^{K} p_k \prod_{i=1}^{n_j} f(y_{ij} \mid LP_{ijk})$$

and the respective log likelihood function is obtained as

$$LL(\beta_1, \beta_2, p, b, z) = \sum_{j=1}^{J} \log \sum_{k=1}^{K} p_k \prod_{i=1}^{n_j} f(y_{ij} \mid LP_{ijk}),$$

where

$$LP_{ijk} = \beta_1^T X_{ij} + \beta_2^T X_j + z_k + b_k x_{ij}$$

or

$$LP_{ijk} = \beta_1^T X_{ij} + \beta_2^T X_j + z_k + b_k x_{ij} + \log(N_{ij})$$

for the binomial model and for the Poisson model, respectively. These are the likelihood function and the log likelihood function of a finite mixed generalized linear model. An EM-algorithm and respective GLIM programs to compute the ML-estimation of such models are described in Dietz and Böhning (1994, 1995) as well as in Aitkin (1999b).

Note that K is an unknown parameter of the log likelihood function above. To find the ML-estimate of K, we maximize LL for a fixed sufficiently large value K. Next, we systematically reduce the value of K to that value K^-, where the maximum of LL decreases for the first time. Then, we consider $\hat{K} = K^- + 1$ as the ML-estimate and the respective estimates of $p = (p_1, p_2, \ldots, p_{\hat{K}})$, $z = (z_1, z_2, \ldots, z_{\hat{K}})$, and $b = (b_1, b_2, \ldots, b_{\hat{K}})$ as the nonparametric estimate of ϕ. An estimate of the mean effect can be obtained by

$$\hat{m}^* = \sum_{k=1}^{\hat{K}} \hat{p}_k \hat{b}_k$$

and its variance by

$$\hat{\tau}^2 = \sum_{k=1}^{\hat{K}} \hat{p}_k \hat{b}_k^2 - \left(\sum_{k=1}^{\hat{K}} \hat{p}_k \hat{b}_k \right)^2 .$$

The posterior probability that the jth study comes from the kth mixture component (C) can be computed by

$$pr(j \in C_k \mid y_{1j}, y_{2j}, \cdots, y_{n_j j}, \hat{p}, \hat{z}, \hat{b}, \hat{\beta}_1, \hat{\beta}_2) = \frac{\hat{p}_k \prod_{i=1}^{n_j} f(y_{ij} \mid \widehat{LP}_{ijk})}{\sum_{r=1}^{\hat{K}} \hat{p}_r \prod_{i=1}^{n_j} f(y_{ij} \mid \widehat{LP}_{ijr})}, \quad (12.4)$$

where

$$\widehat{LP}_{ijr} = \hat{\beta}_1^T X_{ij} + \hat{\beta}_2^T X_j + \hat{z}_r + \hat{b}_r x_{ij} + \log(N_{ij})$$

for the Poisson models and without the last term for the binomial models. These probabilities can be used to obtain a classification of the studies. Such a classification is useful not only for a description of the heterogeneity but also for identification of further explanations of the heterogeneity in addition to the explanatory variables in the model. We now illustrate the method on data of two recently published meta-analyses.

12.4 EXAMPLES

12.4.1 Central Venous Catheters

The first example is a meta-analysis to assess the efficacy of chlorhexidine-silver sulfadiazine-impregnated central venous catheters for the prevention of nosocomial catheter colonization (NCC) and catheter-related bloodstream infection, described in Veenstra et al. (1999). We will reanalyze the published NCC data. Table 12.3 contains the count data of the 12 studies included in this meta-analysis. Here, the units are the sets of impregnated catheters and the sets of non-impregnated catheters in the studies. Thus, we have two units per study, totaling 24 units. The outcome variable is the number of catheter colonizations identified by the same culture techniques of intravascular catheter segments.

Table 12.3 Count Data and Characteristics of 12 Studies on the Efficacy of Chlorhexidine-Silver Sulfadiazine-Impregnation of Central Venous Catheters for the Prevention of Nosocomial Catheter Colonization

	Impregnated			Non-Impregnated			Study Characteristics		
Study	n_1	N_1	MCD	n_0	N_0	MCD	YEAR	CEX	PP
1	8	137	5.1	32	145	5.3	1997	0	0
2	28	208	6.0	47	195	6.0	1997	1	0
3	4	28	6.6	10	26	6.8	1996	0	0
4	22	68	7.0	22	60	8.0	1996	-	0
5	0	14	7.0	4	12	7.0	1994	0	0
6	2	116	7.7	16	117	7.7	1996	0	2
7	60	151	8.5	82	157	9.0	1988	1	0
8	2	98	9.0	25	139	7.3	1999	1	3
9	15	124	9.6	21	127	9.1	1996	1	0
10	45	199	10.9	63	189	10.9	1994	0	0
11	16	123	11.2	24	99	6.7	1995	1	0
12	10	44	8.0	25	35	7.6	1997	1	1

Note. n_1 = number of colonized impregnated catheters, N_1 = number of impregnated catheters, MCD = mean catheter duration, n_0 = number of colonized non-impregnated catheters, N_0 = number of non-impregnated catheters, YEAR = year of study, CEX = catheter exchange, PP = patient population.

Besides the count data, several characteristics of the studies and of the units were obtained and could be used as first and second level variables in our analysis. Table 12.3 contains one first level variable, which is the mean catheter duration (MCD) in the unit, and three second level variables, which are "year of the study", a binary variable, which indicates whether catheter exchange took place within the study (CEX), and a 4-categorical characteristic of the patient population of the study (PP). Their categories are: 1 = transplant ward, 2 = surgical ward, 3 = emergency department, and 0 = other wards. The count data in Table 12.3 yield prevalence rates. Therefore, it is reasonable to use the odds ratio as measure of efficacy. Application of the standard method described in Section 12.1.3 yields an estimate of a common odds ratio $OR = 0.47$ (0.38, 0.57), where the numbers within the brackets are the lower and the upper bound of its 95% confidence interval.

Since the heterogeneity of the 12 studies is highly significant ($Q = 26.7$, $p = .005$), a common efficacy of all studies cannot really be assumed. Therefore, confidence intervals cannot be interpreted. One has to switch to a random effect model. After computing the effect heterogeneity as $\tau^2 = 0.202$ on the basis of the standard random effect model assumptions, the odds ratio estimate $OR^* = 0.39$ (0.27, 0.55) as a mean effect estimate is obtained. This estimate indicates an even higher effect of the chlorhexidine-silver sulfadiazine-impregnation than the common effect size estimate because its value and also its upper confidence bound is lower.

Nevertheless, the standard random effect estimate is doubtful because its assumption of a normal distributed random effect cannot be verified by the data available. If this assumption is not true, and if nothing is known about the real distribution of the odds ratios, then few practical conclusions can be drawn from this result. Generally, it holds that a mean prevention effect does not contradict the possibility that the prevention measure actually increases the infection risk in some of the studies. So, a general recommendation of the measure cannot be given.

To overcome the drawback of the standard method, the nonparametric maximum likelihood approach described in Sections 12.2 and 12.3 is used. In a first step, mixed logistic regression models with a fixed treatment effect ($b_k = m, \forall k$) and without further covariables were fitted. The intercept was the only random effect in these models. We call this step "analysis of baseline heterogeneity". We started with $K = 8$ mixture components and reduced this number systematically. The first increase of the deviance (decrease of log likelihood) can be observed from $K = 4$ (deviance = 69.4) to $K = 3$ (deviance = 78.3). So 4 was considered as the maximum likelihood estimate of K. The mean treatment effect estimate in this 4-component mixture model is $OR^* = 0.44$ (0.33, 0.59).

In a second step, called "analysis of effect heterogeneity", the fixed treatment effect in the model is replaced by a random effect. Again, the respective nonparametric maximum likelihood estimate of this model can be obtained by the maximum likelihood estimation of a finite mixture model. As an estimate of the number of components, $\hat{K} = 5$ with a deviance 49.1 is obtained. On the basis of the 5-component mixture model, the mean effect size estimate is $OR^* = 0.33$ (0.22, 0.49). Note that, although this approach uses weaker model assumptions, its effect estimator indicates a slightly stronger treatment effect than those obtained as fixed model estimate and as standard random effect estimate. However, as already mentioned above, it is difficult to interpret a mean effect size if nothing is known about the distribution of the random effect. One nice property of our approach is that we have an estimate of the whole random effect distribution as a byproduct. This is a finite mixture distribution with 5 components in this case. Each component has its own effect size estimate and each study can be classified into one of these components by Equation 12.4. The results are shown in Table 12.4.

Each of the 5 component effect size estimates are smaller than one, although they are not statistically significant in the third and fourth component. Thus, our meta-analysis provides some evidence that the use of chlorhexidine-silver sulfadiazine-impregnation generally reduces the risk of catheter colonization.

There are two mixture components (1 and 5) with an especially high treatment efficacy. Their odds ratio estimates are $OR_1 = 0.1$ and $OR_5 = 0.11$, respectively, whereas the odds ratios of the other components are about 0.5. Three studies (6, 8, 12) are allocated to these components. In order to describe

Table 12.4 Nonparametric ML-Estimate of the Treatment Effect Distribution and Classification of the Studies by Their Posteriori Component Membership Probability

Component (k)	$OR_k = \exp(b_k)$	CI	p_k	Allocated Studies
1	0.10	(0.03, 0.42)	.19	6, 8
2	0.44	(0.28, 0.69)	.42	1, 2, 5, 9, 11
3	0.62	(0.36, 1.04)	.21	3, 4, 10
4	0.60	(0.32, 1.14)	.09	7
5	0.11	(0.03, 0.50)	.08	12

Note. OR_k = odds ratio as effect size estimate, CI = confidence interval, p_k = posteriori component membership probability.

situations where the efficacy of the prevention measure is particularly high, one should try to characterize their study population. Study 12 is the only study which was exclusively performed in a transplant ward. The patients of the studies 6 and 8 were from a surgical ward and an emergency department, respectively. Thus, if the patient population of the studies could be considered as a representative sample of the all patients in the respective wards, then there would be some evidence that the chlorhexidine-silver sulfadiazine-impregnation should be recommended especially in transplant, surgical, and emergency wards. In order to obtain a more complete picture, we tried to explain the two kinds of heterogeneity in the studies not only by the variable "patient population" but also by all covariables available. Only two second level variables turned out to provide some significant explanation. These are the mean catheter duration in the control group ($MCD0$) and the binary indicator of the transplant ward (TU). In order to explain not only the baseline heterogeneity but also the effect heterogeneity, we additionally included the interaction terms $TU \cdot x_{ij}$ and $\Delta MCD \cdot x_{ij}$ into the model, in which x_{ij} is the treatment indicator and ΔMCD denotes the difference of the mean catheter duration between control group and treatment group. The latter quantity is a derived second level variable which serves to adjust the effect estimate for the potential bias introduced by non-zero differences.

The inclusion of these variables explains a great deal of the heterogeneity in the data. The respective nonparametric ML-estimate of the random effect model is only a two-component mixture. Table 12.5 shows the effect estimates of the covariables in this model.

Both TU and its interaction with the treatment are significant. Consequently, the transplant ward study is a main source of both baseline heterogeneity and effect heterogeneity.

Also, the baseline mean catheter duration has a significant positive effect. That is, the larger the catheter duration, the higher the risk of catheter colonization. The variable $MCD0$ can only explain baseline heterogeneity. The

Table 12.5 ML-Estimates of the Coefficients of the Covariables in a 2-Component Mixed Logistic Regression Model

Variable	Estimate	Standard Error
TU	1.156	0.531
$MCD0$	0.134	0.040
$TU \cdot x_{ij}$	-1.698	0.722
$\Delta MCD \cdot x_{ij}$	0.010	0.086

Note. TU = binary indicator of the transplant ward, $MCD0$ = mean catheter duration in the control group, x_{ij} = treatment indicator, ΔMCD = difference of the mean catheter duration between control group and treatment group.

term $\Delta MCD \cdot x_{ij}$ is not significant and provides only inconsiderable explanation of effect heterogeneity.

Table 12.6 shows the nonparametric ML-estimate of the treatment effect and the respective classification of the studies. Now, most of the studies are classified into one component with an adjusted effect estimate of $OR = 0.43$ (0.30, 0.62). Thus, some evidence has been obtained to recommend the prevention measure in the patient population of all studies in this component. The high efficacy of the catheter impregnation in the study 12 is already shown by the significance of the term $TU \cdot x_{ij}$ of the model.

Table 12.6 Nonparametric ML-Estimate of Treatment Effect Distribution Adjusted for Covariables and Classification of the Studies by Their Posteriori Component Membership Probability

Component (k)	$OR_k = \exp(b_k)$	CI	p_k	Allocated Studies
1	0.43	(0.30, 0.62)	.73	1, 2, 3, 5, 6, 8, 9, 10, 11
2	0.64	(0.37, 1.10)	.27	4, 7, 12

Note. OR_k = odds ratio as effect size estimate, CI = confidence interval, p_k = posteriori component membership probability.

The only populations where the efficacy of the prevention measure remains questionable are those of studies 4 and 7, which are allocated to the second component. The strategy of further research could be to look for specific characteristics of these two studies, which can explain their worse results.

12.4.2 Ischaemic Heart Disease Events

The data of the second example are given by Thompson and Sharp (1999). They are taken from 28 randomized trials in which ischaemic heart disease events are considered as a response variable. An ischaemic heart disease event is defined as a fatal ischaemic heart disease or a non-fatal myocardial infarction. Another response variable of these trials is the average serum cholesterol reduction. However, this variable will be used as an explanatory variable of the effect and baseline heterogeneity.

This meta-analysis is not a very typical one in hospital and clinical epidemiology because the prevention measures studied within the trials are very different. The spectrum of applied prevention measures includes dietary interventions, drugs, and even surgery. However, this meta-analysis is very suitable to demonstrate certain potentials of our method. It will be shown that a meta-analysis can be accomplished sensibly even when the studies considered have different study factors.

Trial-specific count data and cholesterol reductions are given in Table 12.7. Also, the count data in Table 12.7 are prevalence-type data. Therefore, the odds ratio is used again as measure of the efficacy. By applying the standard method, an estimate of a common odds ratio $OR = 0.82$ (0.77, 0.88) can be obtained.

The heterogeneity of the 28 studies is highly significant ($Q = 49.1$, $p = .006$), which is expected because of the different study factors of the studies. The estimated effect variance is $\tau^2 = 0.202$ and the respective mean odds ratio estimate based on the standard random effect model is $OR^* = 0.81$ (0.72, 0.90).

Now, one could proceed as in the previous example and estimate the random effects distribution. The aim of this study is not to describe the random effect distribution but to explain the variation of the odds ratios by the variable "mean cholesterol reduction". Therefore, we computed the nonparametric ML-estimation of a random effect model with the explanatory variables "study" and "mean cholesterol reduction" and with a random treatment effect.

By assuming that the value of the mean serum cholesterol reduction is equal to zero in the control groups, this variable can be considered a first level variable. The categorical variable "study" provides the complete explanation of baseline heterogeneity. $\hat{K} = 1$ is obtained as estimate of the number of mixture components. Consequently, the respective table of the effect distribution estimate and of the study classification has one line only (see Table 12.8).

Thus, the whole effect heterogeneity is explained by the variable mean cholesterol reduction. Its estimated logistic regression coefficient is -0.479 (0.14), where the number within the brackets is the respective standard error. The adjusted common treatment effect estimate is $\ln(OR) = 0.122$ (0.10). It is not significant. The following conclusions can be drawn from these results now:

1 The heterogeneity of the effect sizes in this meta-analysis can be explained by the variable "mean cholesterol reduction".

2 The significant mean treatment effect size estimate can be explained by the cholesterol reduction attained by the prevention measures.

3 The heterogeneity of the effect size can be explained by the heterogenous effects of the several prevention measures on cholesterol reduction.

4 Serum cholesterol reduction should be a main goal for ischaemic heart disease prevention.

Table 12.7 Count Data and one Study Characteristic of 28 Clinical Trials on the Efficacy of Diverse Prevention Measures to Reduce the Risk of Ischaemic Heart Disease

	Control group		Treatment Group		Cholesterol
Trial	n_0	$N_0 - n_0$	n_1	$N_1 - n_1$	Reduction(mmol/l)
1	210	5086	173	5158	0.55
2	85	168	54	190	0.68
3	75	292	54	296	0.85
4	936	1853	676	1546	0.55
5	69	215	42	103	0.59
6	101	175	73	206	0.84
7	193	1707	157	1749	0.65
8	11	61	6	65	0.85
9	42	1087	36	1113	0.49
10	2	28	2	86	0.68
11	84	1946	56	1995	0.69
12	5	89	1	93	1.35
13	121	4395	131	4410	0.70
14	65	357	52	372	0.87
15	52	142	45	154	0.95
16	81	148	61	168	1.13
17	24	213	37	184	0.31
18	11	41	8	20	0.61
19	50	84	47	83	0.57
20	125	292	82	339	1.43
21	20	1643	62	6520	1.08
22	0.5	52.5	2	92	1.48
23	0.5	29.5	1	22	0.56
24	5	25	3	57	1.06
25	144	871	132	886	0.26
26	24	293	35	276	0.76
27	4	74	3	76	0.54
28	19	60	7	69	0.68

Note. IHD events = fatal ischaemic heart disease and non-fatal myocardial infarction, n_0 = number of patients with an IHD event in the control group, N_0 = number of patients in the control group, n_1 = number of patients with IHD event in the treatment group, N_1 = number of patients in the treatment group.

Table 12.8 Nonparametric ML-Estimate of Treatment Effect Distribution Adjusted for Covariables and Classification of the Studies by Their Posteriori Component Membership Probability

Component (k)	$OR_k = \exp(b_k)$	CI	p_k	Allocated studies
1	1.13	(0.92, 1.37)	1.00	all studies

Note. OR_k = odds ratio as effect size estimate, CI = confidence interval, p_k = posteriori component membership probability.

12.5 CONCLUSION

Baseline and effect heterogeneity are almost always present in hospital and clinical epidemiological meta-analyses. This makes the evaluation of a treatment effect difficult and the use of some standard methods questionable. Attempts to restrict the meta-analysis on a homogenous selection of studies can never be completely successful. On the other hand, such attempts usually mean renouncing valuable information. It is now generally agreed that meta-analysis can and should go further than simply producing overall summaries of effects. In particular, understanding the possible causes of any heterogeneity can increase both the scientific value and clinical and epidemiological relevance of the results from a meta-analysis. In this chapter, an appropriate method for addressing this issue is presented. This method is based on finite mixed generalized linear models (FMGLMs), which have proven to be very flexible tools to estimate mean effect sizes and explain heterogeneity. Different effect measurements can be considered simply by changing the link function of the model. For example, using the identity link instead of the logit link leads to meta-analyses of risk differences instead of odds ratios.

The analysis of the example data shows that much more information can be gained by this approach than by the standard method. The second example clearly shows that the heterogeneity itself can be the most interesting subject of analysis.

There are of course some disadvantages and limitations to this approach. The first limitation is the fact that the count data of the studies must be available, which is not always the case. A second limitation is that the number of studies in the meta-analysis should be larger then 10. Otherwise, the nonparametric maximum likelihood estimation will be doubtful. A third limitation concerns the number of covariables. This number should not be too large in comparison to the number of studies. Finally, it should be noted that misinterpretation of the effect of second level variables is possible, especially if they are mean values of the study population.

REFERENCES

Aitkin, M. (1999a). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117–128.

Aitkin, M. (1999b). Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, *18*, 2343–2351.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping, and others.* Boca Raton, FL: Chapman & Hall/CRC.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Dietz, E., & Böhning, D. (1994). Analysis of longitudinal data using a finite mixture model. *Statistical Papers*, *35*, 203–210.

Dietz, E., & Böhning, D. (1995). Statistical inference based on an general model of unobserved heterogeneity. *Lecture Notes in Statistics*, *104*, 75–82.

Olkin, I. (1999). Diagnostic statistical procedures in medical meta-analysis. *Statistics in Medicine*, *18*, 2231–2341.

Thompson, S. G. (1993). Controversies in meta-analysis: The case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research*, *2*, 173–192.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*, 2693–2708.

Veenstra, D. L., Saint, S., Saha, S., Lumley, T., & Sullivan, S. D. (1999). Efficacy of antiseptic-impregnated central venous catheters in preventing catheter-related bloodstream infection: A meta-analysis. *Journal of the American Mathematical Association*, *281*, 261–267.

13

A Generalized Linear Model Incorporating Measurement Error and Heterogeneity Applied to Meta-Analysis of Published Results in Hodgkin's Disease

Jeremy Franklin

Biometrics Group
First Department of Internal Medicine
University of Cologne

Summary

For a meta-analysis based on comparisons between studies rather than controlled comparisons within studies, it is especially important to explain, estimate, and to allow for the heterogeneity in results between studies. The danger of bias in the use of "historical controls" is well known. The aim of the present investigation was to develop a simple method to analyze differences in results between paediatric and adult clinical trials in Hodgkin's disease, through meta-analysis of a full and extensive collection of published results. Patient and treatment characteristics were included as possible explanatory factors in a generalized linear model. Sampling errors in the Kaplan-Meier estimates derived in the studies as well as heterogeneity between studies were estimated iteratively in order correctly to weight the observations and assess significance while fitting the model and testing effects. A significant superiority of treatment results in children, compared with adults, was demonstrated, allowing for patient and treatment characteristics. A generalized linear model incorporating heterogeneity and explanatory factors was found to be a practicable and flexible method for a "between-studies" meta-analysis, suitable for investigations where controlled comparisons are not possible or not available.

13.1 INTRODUCTION

The "standard" type of meta-analysis combines the results of several randomized studies, each of which makes the same (randomized) comparison as the meta-analysis. The technique has been extended to non-randomized investigations, for example, of diagnostic methods or prognostic factors (see Chapters 11 and 6 in this volume). Again, such meta-analyses combine comparative information (relative frequencies, correlations, etc.) from each study. The meta-analysis addresses the same question as each component study, its advantage lying in the greater (combined) sample size and therefore power, and in its greater representativity. In practice, many questions and hypotheses in clinical research have not been – or cannot be – investigated within studies. Tentative inferences are then made using comparisons *between* studies, for instance, a historical comparison between a former treatment and a new treatment. Such comparisons are liable to suffer from hidden or unquantifiable biases due to systematic differences between the patients, or their treatment, in the compared studies. Two techniques may help to improve the reliability of between-study comparisons: firstly, avoiding selection bias and "averaging-out" of chance differences by systematic inclusion of a large number of relevant studies; secondly, modeling the influence of known study characteristics to make allowance for biases. This report applies these techniques to the comparison of treatment results in Hodgkin's disease between paediatric and adult institutions.

13.2 OBJECTIVE

In the development and optimization of therapy for malignant lymphoma it has been widely observed during the last decades that paediatric treatment results, as a whole, are superior to those achieved in adults. Since there are systematic differences in treatment strategy between paediatric and non-paediatric institutions (emphasis on combined chemo-radiotherapy even for early stages, lower radiation doses, new and more intensive chemotherapies for children), the reason for this superiority and the role of patient age were unclear. Should the type of therapy rather than age per se be responsible for the good paediatric results, then a rethinking of adult therapy and a borrowing of ideas from children's institutions might be fruitful (Magrath, 1997). The aim of the present analysis was to model the dependence of cure rates in paediatric and non-paediatric Hodgkin's disease patient cohorts on the factors age-range of patients, distribution of disease characteristics in the cohort and type of therapy. The size of differences, if any, in cure rates between children's and adults' cohorts with similar disease characteristics and therapy were to be evaluated.

13.3 METHODS

The publications of studies from which data were to be extracted were selected using a systematic search in the medical literature database Medline (1980–1997) followed by the application of several predefined criteria (first-line treatment, sample size at least 30, chemotherapy as main therapy component, adequate information and follow-up). Pure radiotherapy trials were omitted since children are rarely treated with radiation alone. The patients reported in each paper were, where appropriate and as far as the sample size and the reported details allowed, divided into homogeneous cohorts according to disease characteristics and therapy, avoiding subgroups of less than 30 patients.

Data concerning the type of institution or trial group, sample size, distribution of disease characteristics, type of therapy and Kaplan-Meier estimates of cure rates (disease free survival (DFS) and overall survival (SV) rates) were extracted. Cure rates were adjusted to the time point 5 years after first diagnosis, this adjustment being based on results of a linear regression on pooled data at multiple time points from all those publications where a Kaplan-Meier plot covering an adequate time span was given.

The form of this meta-analysis was a generalized linear model (McCullagh & Nelder, 1989; for some further developments see Nelder, 1998) with cure rate (S) as response variable:

$$\log\left(-\log(E(S))\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

In order to restrict predicted rates to between 0 and 100%, the complementary log-log link function was chosen to relate response variable S to the linear combination of explanatory variables X and unknown parameters β. The observed response is assumed to vary normally about the expected value with variance comprising two components, namely the sampling error of the Kaplan-Meier estimate σ_P^2 (P = variation between patients) and the heterogeneity σ_C^2 (C = variation between cohorts):

$$S \sim \mathcal{N}\left(E(S);\, \sigma_P^2 + \sigma_C^2\right).$$

Nine potentially explanatory variables were considered, namely:

- type of institution (single-centre, oligocentric, multicentric)
- recruitment period
- sample size
- proportion with advanced disease (stages III and IV)
- proportion with systemic symptoms
- treatment modality (chemotherapy or combined chemo-radiotherapy)

- type of chemotherapy
- number of different drugs
- number of cycles of chemotherapy.

The observations were weighted according to their estimated variance: the standard error of the (iteratively estimated) cure rate according to the formula of Greenwood (1926) plus a second component allowing for the heterogeneity between cohorts. The amount of heterogeneity was estimated iteratively from the sum of squared model residuals, allowing for the contribution due to σ_P^2 (iterative reweighting).

Model fitting was performed in SAS by iterative use of the procedure GENMOD. After each model fitting, the fitted values of cure rate and the residuals were used to estimate the Greenwood standard errors and the heterogeneity respectively. These estimates were combined to estimate the variance of each observation and hence its weight for the next fit, until (after typically 4–5 iterations) the results were stable (Figure 13.1). Using stepwise regression techniques, the effect of including or excluding each explanatory factor was assessed by the change in log-likelihood. Thus, an "optimal" model including only the significant explanatory factors was selected.

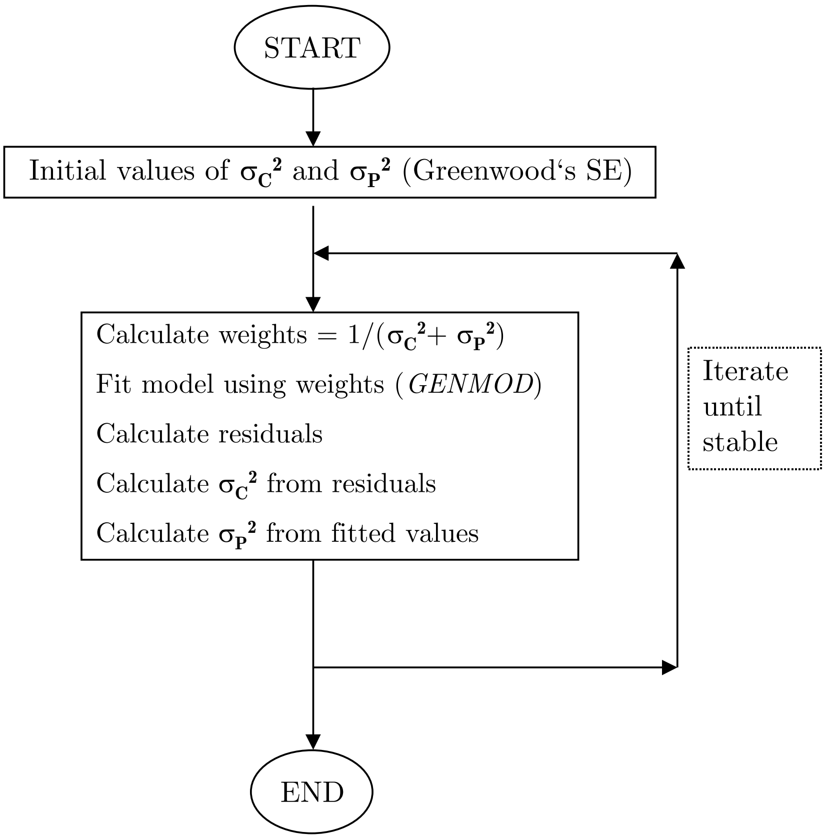

START
Initial values of σ_C^2 and σ_P^2 (Greenwood's SE)
Calculate weights = $1/(\sigma_C^2 + \sigma_P^2)$
Fit model using weights (*GENMOD*)
Calculate residuals
Calculate σ_C^2 from residuals
Calculate σ_P^2 from fitted values
Iterate until stable
END

Figure 13.1 Iteratively reweighted fitting of the generalized linear model using the SAS procedure GENMOD.

The sensitivity of the results to small changes in the model (choice of explanatory factors, logistic link function, uniform weighting, weighting proportional to sample size) was investigated. The model was also fitted to several subgroups of the available cohorts (combined modality treatments, pure chemotherapy treatments, particular chemotherapy schemes, early stages, advanced stages, larger cohorts) to assess the generalizability of the results.

In order to correct for the poorer prognosis of the elderly patients who were represented in almost all adult cohorts, the age-specific cure rates of patients in the multicentre German Hodgkin's Lymphoma Study group were analyzed. The effect of older patients in lowering the cure rate of the whole cohort was estimated. Allowing for this effect enabled us to use the meta-analysis results to compare children with younger adults alone.

13.4 RESULTS

Thirty-eight paediatric and 85 adult cohorts were selected for inclusion. The distribution of disease characteristics was similar, on average, in paediatric and non-paediatric cohorts. However, consistent differences in type of cohort and type of therapy were seen (Tables 13.1 and 13.2). Due to the lower incidence of Hodgkin's disease in children, the paediatric cohorts were on average smaller, more often multicentric, and less often randomized. Children more often received combined modality therapy and a lower radiation dose, and therapy more often included a more modern, ABVD[1]-type regimen.

The heterogeneity of both DFS and SV rates between cohorts was significant, according to the *Q*-statistic of DerSimonian and Laird (1986). For DFS, heterogeneity represented about two-thirds of the residual variation and for SV, about one-third.

Cure rates were consistently better, on average, for paediatric cohorts than for adults. Figures 13.2 – 13.4 show examples of the distribution of paediatric and adult DFS plotted against three of the potential explanatory factors, allowing paediatric and adult results (with respect to the chosen factor) to be compared. This graphical method of comparison is limited to single explanatory factors.

Using the generalized linear modeling technique with all the explanatory factors listed above, highly significant differences of circa 13% in DFS (95% confidence interval: 6–19%) and 12% in survival (95% confidence interval: 8–15%) were found. These differences in cure rates were calculated from the estimated regression coefficients for the factor paediatric/adult via the link function (see above). Figure 13.5 shows estimated differences, which (due to the curvature of the link function) vary according to the level at which the cure rates lie. Alongside the factor paediatric/adult, the following factors were selected as significant for DFS by the stepwise fitting procedure: type of institu-

[1] Adriamycin, Bleomycin, Vinblastine, Dacarbazine

Table 13.1 Characteristics of Paediatric and Adult Cohorts

		Pediatric		Adult	
Type of Cohort	Single Centre	14	37%	50	59%
	2-4 Centres	–	–	7	8%
	Multicentre	24	63%	28	33%
Randomized Study?	No	29	76%	40	47%
	Yes	9	24%	45	53%
Number of Patients	< 40	9	24%	15	18%
	41-60	11	29%	19	22%
	61-100	15	40%	27	32%
	101-200	3	8%	17	20%
	> 200	–	–	7	8%
Total		38		85	

Note. Entries represent number and percentage of cohorts.

Table 13.2 Treatment in Paediatric and Adult Cohorts

		Pediatric		Adult	
Chemotherapy	MOPP or similar	8	21%	42	49%
	ABVD or similar	7	18%	5	6%
	MOPP/ABVD or similar	9	24%	13	15%
	OPPA	9	24%	–	–
	Other	5	13%	25	29%
Number of Chemotherapy Cycles	2–3	7	20%	17	21%
	4	6	17%	5	6%
	5–7	16	44%	40	49%
	8	5	14%	14	17%
	> 8	2	6%	6	7%
Radiotherapy Fields	None	5	13%	26	31%
	Localized	27	71%	18	21%
	Extended	6	16%	33	39%
	Various	–	–	7	8%
Total		38		85	

Note. Entries represent number and percentage of cohorts. MOPP = Mustargen, Vincristine, Procarbacine, Prednisone, ABVD = Adriamycin, Bleomycin, Vinblastine, Dacarbazine, OPPA = Vincristine, Procarbacine, Prednisone, Adriamycin.

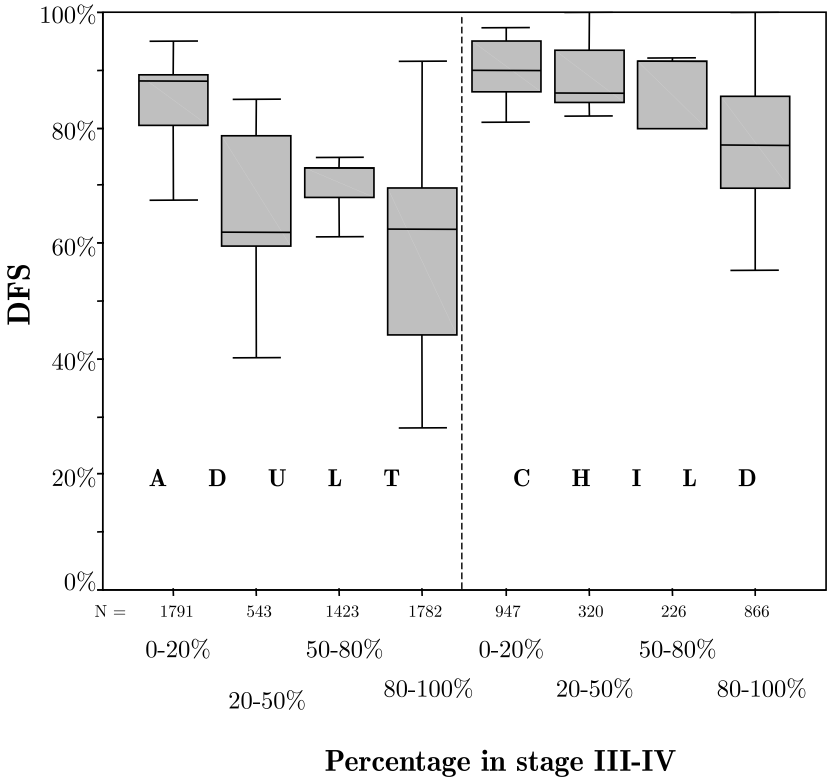


Figure 13.2 Boxplots of (5-year-adjusted) disease free survival rates according to the proportion of advanced stage patients in the cohort, for adult (left) and paediatric (right) cohorts.

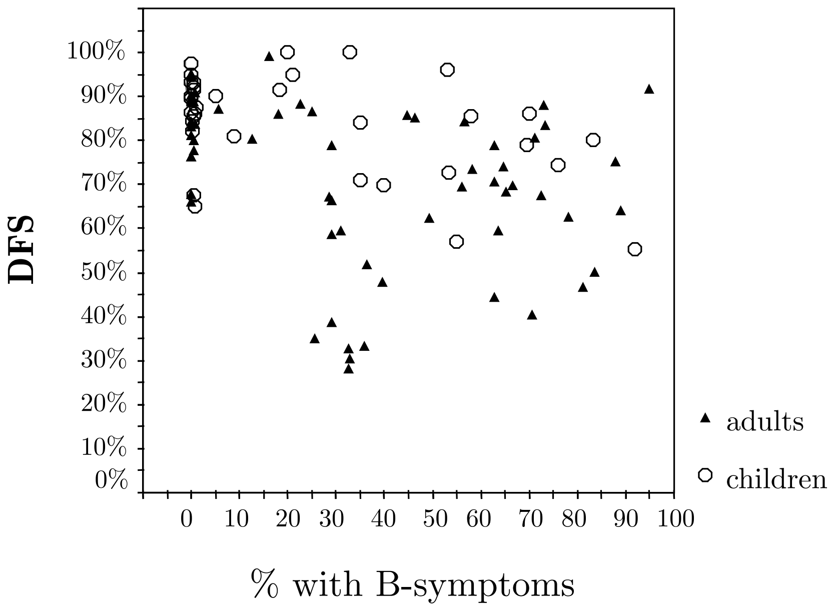


Figure 13.3 Scatterplot of disease free survival (5-year adjusted) against proportion of patients with systemic (B) symptoms in the cohort, for paediatric and adult cohorts.

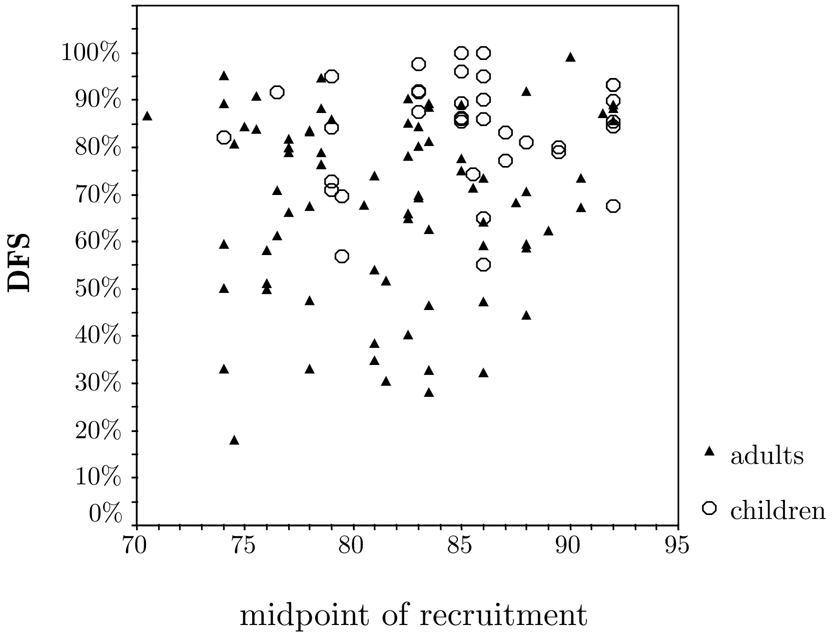


Figure 13.4 Scatterplot of disease free survival (5-year adjusted) against year at midpoint of recruitment period, for paediatric and adult cohorts.

tion, proportion with disease stage III-IV, proportion with systemic symptoms, treatment modality).

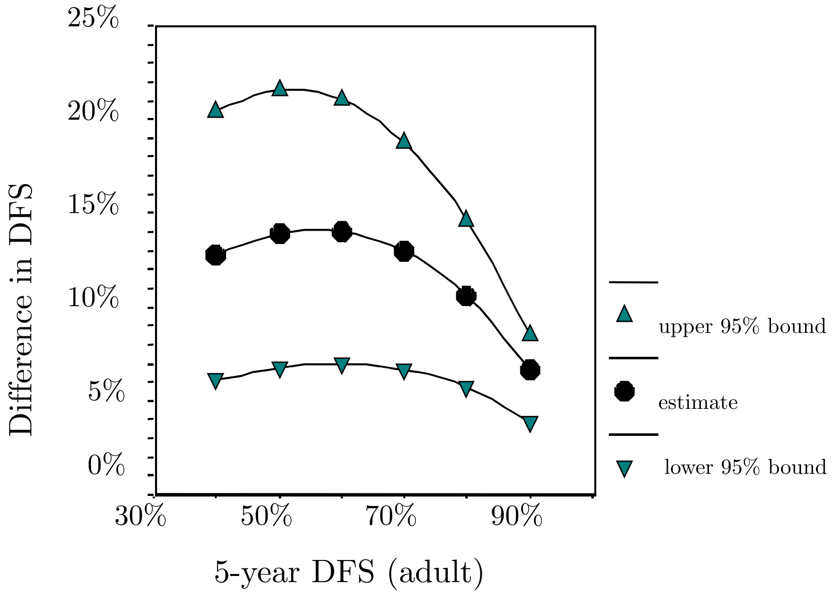


Figure 13.5 Estimated differences in disease free survival rates between paediatric and adult cohorts, calculated using the estimated parameter $\beta = 0.442$ from the generalized linear model.

The paediatric/adult difference was not restricted to certain types of cohorts but reappeared as significant in all main subgroups of cohorts, including the cohorts of predominantly early stage patients, the cohorts of predominantly advanced stage patients, the cohorts receiving combined chemo-radiotherapy, the cohorts of size over 80, the multicentre cohorts, and so forth. Small variations in modeling methods (weighting scheme, form of link function) did not qualitatively change the results.

The reduction in DFS and SV due to the presence of patients over 45 years old in the adult cohorts was estimated as 3% and 4% respectively. The remainder of 9% in each case therefore represents a difference between children and young adults. This difference could not be accounted for by therapy-related or other factors. It could be due to an intrinsic biological difference or to hidden confounding factors such as quality of care in paediatric institutions.

13.5 CONCLUSIONS

A statistical model for the dependence of treatment results on explanatory factors relating to treating institution, patient cohort and type of therapy was constructed and fitted, with the aim of estimating the difference in treatment results attributable to the age range of the patients (paediatric or adult, respectively).

The generalized linear model allows an appropriate form of dependence and an appropriate specification of error to be incorporated. Heterogeneity between cohorts was an important part of the random variation in both endpoints. The iterative estimation of heterogeneity together with an approximate calculation of the standard error of each cohort-based Kaplan-Meier estimate leads to an error structure which allows for both types of error and thus to a plausible weighting scheme. In the application to Hodgkin's disease, the size of the effect of interest (superiority of paediatric cure rates) could be estimated, although the precision was not high. The results were not sensitive to small changes in modeling methods or inclusion criteria.

A more sophisticated, integrated approach would make use of a maximum likelihood technique to fit a generalized linear model with variance components, the heterogeneity between cohorts being represented by a cohort random effect. Aitkin (1999) lists alternative techniques and proposes a nonparametric approach.

The presence of hidden or non-quantifiable differences which occur systematically between paediatric and adult *cohorts* could influence the results of such an analysis despite the attempt to explain and allow for heterogeneity. The possible influence of such effects should be carefully assessed. In the present analysis, intrinsic differences in curability between children and adults may be confounded with different treatment strategies adopted by paediatric compared with adult institutions. The inclusion of treatment type as a factor in the model may only partially allow for such confounding. Furthermore, it has been suggested that children systematically receive a more thorough staging, treatment administration and care, factors which are not available for inclusion in the model. Thus, the conclusion that paediatric results are superior, for comparable patient characteristics and treatment schemes, applies only under the current standards of care and management in paediatric and adult institutions respectively.

The credibility of the results of any meta-analysis, but especially one based on comparisons *between* studies, depends on the unbiased and representative selection of studies for inclusion. In the present analysis, credibility was sought through systematic inclusion of a large number of studies.

REFERENCES

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117–128.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Greenwood, M. (1926). The errors of sampling in the survivorship tables (Appendix 1). In *Reports on public health and statistical subjects.* London: HMSO.

Magrath, I. T. (1997). The treatment of pediatric lymphomas: Paradigms to plagiarize? (Supplement 1). *Annals of Oncology*, *8*, 7–14.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Nelder, J. A. (1998). A large class of models derived from generalized linear models. *Statistics in Medicine*, *17*, 2747–2753.

Acknowledgments

The author is grateful to the chairmen of the German paediatric and adult Hodgkin's disease study groups, respectively Professors G. Schellong and V. Diehl, for motivating the work, and to several colleagues in Cologne for fruitful discussions.

14

The Influence of Design Variables on the Results of Controlled Clinical Trials on Antidepressants: A Meta-Analysis

Claudia Schöchlin
Jürgen Klein
Department of Clinical Psychology and Psychophysiology
Hospital for Psychiatry and Psychotherapy
Ludwig-Maximilians-University of Munich

Dorothee Abraham-Rudolf
Hospital for Psychiatry
Technical University of Munich

Rolf R. Engel
Department of Clinical Psychology and Psychophysiology
Hospital for Psychiatry and Psychotherapy
Ludwig-Maximilians-University of Munich

Summary

Bias in clinical trials can be investigated by studying correlations between design variables and outcome of clinical trials. 72 studies on the antidepressant effect of imipramine and amitriptyline as well as four serotonin reuptake inhibitors were analyzed in a publication based meta-analysis. Treatment outcome was operationalized as an effect size in the basis of response rate differences between active drug and placebo or active drug and active drug. It was found that the number of treatment cells included in a study, the existence of a placebo cell as well as the severity of depression at inclusion and placebo response rate are associated with study outcome, and that they may interact with each other, presumably because of differences in drop-out handling. Main conclusion is that the existence and the clinical results of a placebo cell are moderating variables for the results of clinical trials.

14.1 INTRODUCTION

Meta-analysis is a tool to statistically integrate results of empirical studies from publications or, less often, from raw data (Dickersin & Berlin, 1992). Two forms of empirical studies can be distinguished: Experimental or quasi-experimental studies with an independent and a dependent variable, such as clinical trials, and epidemiological studies, such as case-control studies, in which there is no independent variable that is manipulated (Petitti, 1994). Both kinds of study are subject to methodological flaws which, in turn, affect interpretation of meta-analytical results. A critical appraisal of these problems is presented by Feinstein (1995). On the other hand, meta-analysis can serve as a tool to investigate methodological questions arising from empirical studies; an example of this is furnished by Gotzsche (1990), who presents an investigation of methodological problems of trials in rheumatoid arthritis.

The present meta-analysis investigates the effects of design variables of clinical antidepressant trials. Design variables are variables that are voluntarily or involuntarily induced by all aspects concerning the design and the realization of clinical trials. They are not welcome, as they may restrict the ability to generalize study results, and therefore risk impeding the interpretation of data. Design effects can be caused by blindness, number of treatment conditions, handling of placebo-responders, or the baseline severity of illness.

The term “Depression” is frequently used in colloquial language and means sorrow or despair. As a psychiatric diagnosis it is actually operationalized by the persistence of a number of certain, well-defined symptoms (depressive mood, loss of interest, loss of weight, sleep disturbances, psychomotor agitation or retardation, fatigue, cognitive feelings of guilt or worthlessness, disturbances of thought, concentration, memory, decisiveness, suicidality) over a given time period (American Psychiatric Association, 1994). It is one of the most frequent diseases and has large economic consequences for a society as well as, with its infringing symptoms and finally its high suicide rate, important consequences for sufferers’ lives (Goodwin & Jamison, 1990). Therefore, it constitutes one of the main psychiatric research areas, in which a large number of clinical trials has been conducted that need to be meta-analytically integrated and investigated.

Concerning the drugs used in antidepressive therapy, one can distinguish different groups: The first generation drugs were the tricyclic antidepressants (TCA), which are considered as a standard treatment in depression drug therapy. In the 1980s, the selective serotonin reuptake inhibitors (SSRI), which possess more specific receptor activity, were developed. They are considered to be an alternative to their predecessors, as they induce less side effects (Möller & Volz, 1996). Although a very large number of substances has been investigated in clinical trials on drug therapy of depression, most of the studies included TCA or SSRI.

14.2 BACKGROUND

We started from the following result of a former meta-analysis on the tricyclic antidepressant imipramine, compared to placebo: Studies that compared imipramine only to placebo, and not to a third or fourth treatment condition, yielded considerably higher effect sizes ($r = .34$; $N = 703$)[1] than studies which included further treatment arms ($r = .17$; $N = 4673$). Thus, the difference between imipramine and placebo was higher if only imipramine and placebo were investigated ($z = 4.6$; $p < .001$). Greenberg, Bornstein, Greenberg, and Fisher (1992) proposed that a study is less susceptible to unblinding if more than two treatment conditions are investigated. We dared not to join this interpretation as there was a confounding with other variables we considered to have potential weight, that is, the year of publication and the status of a substance as a new or as a control substance. Among the 11 studies published before 1978, only 3 included more than three treatment arms, compared to 60 among the 66 studies published after 1977. Nearly all studies with two treatment conditions are older studies and may therefore be more prone to bias, as methodology in the early years of clinical trials was less elaborate than nowadays. Imipramine was investigated as the substance of interest mainly in studies with only two treatment cells (14 out of 15), whereas it served as a control substance (logically) only in studies with more than two treatment cells. Most of the studies with imipramine as a control substance were published after 1978 (59 out of 62), whereas it was investigated as a substance of interest before and after 1978.

We nevertheless considered the difference between studies with two and those with more than two treatment cells being important and took a further look at it by comparing placebo-controlled SSRI-studies. All studies were published after 1977. There were no publications on placebo-controlled studies with an SSRI as a control substance; all studies investigated the SSRI as the substance of interest. No difference was found among studies with two versus those with more than two treatment arms ($r = .17$, $N = 696$ vs. $r = .18$, $N = 3155$; $z = 0.28$, $p = .78$). Thus, the proposal based on the imipramine results that the number of treatment cells is a decisive factor for the effect size of a study was not affirmed by the SSRI data.

14.3 AIMS OF THE META-ANALYSIS

In the literature, further variables characterizing the design of clinical studies were identified that might influence effect sizes of antidepressant studies.

[1] Here, and in the following text, N always indicates the (total) number of patients.

We investigated correlations between study characteristics and effect size in a larger sample of controlled clinical trials. As outcome, the effect sizes for active medication vs. placebo or standard medication were chosen; they represent differences in efficacy between the two treatment cells. Predictor variables were the number of study centers, a placebo run-in vs. no placebo run-in, the study duration and mean patients' severity of depression at baseline.

14.4 METHODS

Controlled clinical trials on acute treatment of depression with imipramine, amitriptyline, fluoxetine, paroxetine, or sertraline were included if one of these substances was compared either to placebo and/or to one of the substances of the other group. Studies had to be published between January 1979 and April 1997, and had to indicate response rates.

The difference in efficacy between two treatment cells was expressed by the correlation coefficient r, based on the fourfold table φ, which corresponds to the response rate difference and thus can be interpreted as being quite close to clinical practice. Effect sizes were computed on an intent-to-treat-basis; all randomized patients were included in the effect size calculation. It has been discussed which effect size is preferable for clinical trials (Dickersin & Berlin, 1992; Fleiss, 1993), especially odds ratios are commonly used. We consider response rate differences the appropriate measure in our case, for the very reason that the original studies do not use odds ratios but response rate differences, which are comparable to φ in a meta-analysis. Effect sizes were computed by:

$$\varphi = \sqrt{\frac{\chi^2}{N}}.$$

They were weighted for sample size of the studies and averaged via Z-transformation. Homogeneity of study effect sizes was assessed by the usual chi-square test for homogeneity of correlation coefficients (Rosenthal, 1991; Mantel & Haenszel, 1959). Explorative comparisons of specific effect sizes were performed by means of contrasts for group comparisons (for details see Rosenthal, 1991; Rosenthal & Rubin, 1982), according to the following formula:

$$z = \frac{\sum_{j=1}^{k} \lambda_j Z_j}{\sqrt{\sum_{j=1}^{k} (\lambda_j^2 \, (N_j - 3))}}$$

with Z_j being the Fisher-Z-transformed effect size coefficients for the jth of k studies to be compared and λ_j being the orthogonal contrast coefficients summing up to zero. z is the standard normal deviate. For quantitative data, Pearson correlations were used; they were calculated with weights, but significance was judged referring to the number of studies, not the number of patients. Given p values are two-tailed.

Table 14.1 Comparisons Included in the Meta-Analysis

			Year of publication			BL-Ham-D	
	Number of comparisons	*N*	Min	Max	*M*	*M*	*SD*
Drug – Placebo							
Imipramine	46	6176	79	96	87.7	25.0	4.9
Amitriptyline	16	2604	79	95	87.8	25.5	5.1
Fluoxetine	11	1247	85	95	89.4	24.7	3.8
Fluvoxamine	8	1043	83	96	90.9	25.8	2.0
Paroxetine	5	1052	89	93	91.0	28.4	1.5
Sertraline	4	954	90	96	93.8	22.0	6.8
Drug – Drug							
Imipramine-SSRI	25	3380	83	96	90	26.15	3.2
Amitriptyline-SSRI	19	2337	85	96	91	26.09	2.9

Note. *N* = Number of patients; BL-Ham-D = Hamilton at baseline

14.5 DATA

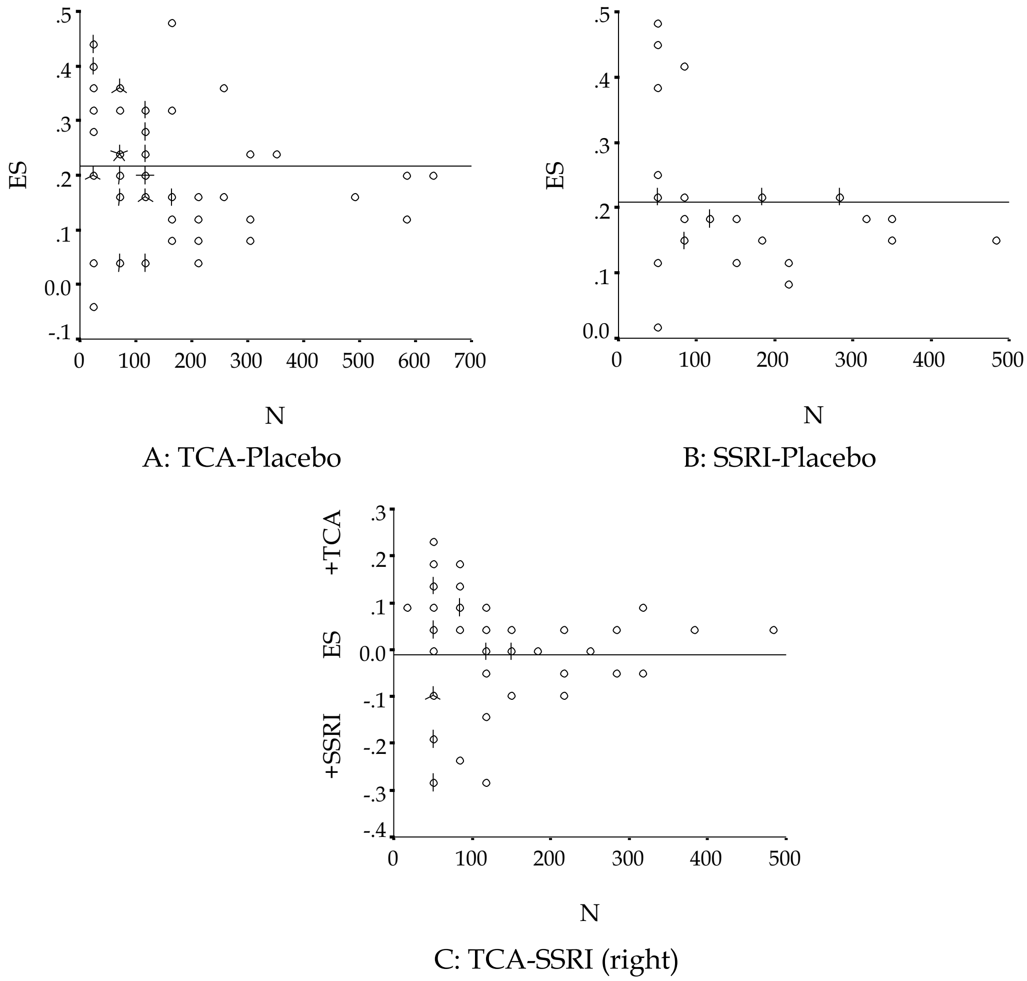
Table 14.1 gives an overview of the comparisons on which the following analysis is based. Sixty-two of the comparisons refer to tricyclic drugs vs. placebo, 28 to SSRI vs. placebo, and 44 to tricyclic drugs vs. SSRI.

14.6 RESULTS

The Funnel plots in Figure 14.1 show that with increasing *sample sizes* effect sizes approach the mean effect sizes. Among the placebo-controlled studies (Figure 14.1, Panel A and B) there are more studies with higher than with lower effect sizes. This was to be expected, as it can be presumed that small positive studies are more likely to be published than small negative studies. This is an indicator of publication bias, which is less clear-cut among the drug-drug-comparisons (Figure 14.1, Panel C).

Data stemming from *single center studies* yielded higher effect sizes than data from multicenter studies (see Table 14.2). This means that active drugs show their superiority to placebo more clearly if the data are collected in only one center. The same effect can be found when comparing SSRI vs. TCA: Multicenter studies show slight differences between substance classes; in single center studies, there is a tendency for the SSRI to yield better results.

If a *placebo run-in* or placebo washout is included in a study, patients receive placebo during one or two weeks before randomization and are excluded if they respond during this time period. This placebo run-in aims at increasing effect sizes by reducing the number of responders of one group, the placebo



A: TCA-Placebo

B: SSRI-Placebo

C: TCA-SSRI (right)

Figure 14.1 Funnel plots (sunflower plot): Number of patients and effect sizes (ES), line represents unweighted mean of effect size).

Table 14.2 Effect Sizes From Single- and Multicenter Studies

	Single-center Studies	Multicenter Studies		
	r	*r*	*z*	*p*(*z*)
TCA-PL	.25 (32)	.18 (30)	2.91	.004
SSRI-PL	.26 (11)	.17 (17)	2.07	.040
TCA-SSRI	−.08 (15)	.01 (28)	−2.51	.010

Note. For each *r*, the number of studies is given in brackets.

group (FDA, 1978; Feinberg, 1992). Only sparse data exist about the effects of this procedure on study results, and these are ambiguous (Khan, Cohen, Dager, Avery, & Dunner, 1989; Trivedi & Rush, 1994). The data presented here revealed no difference between studies with and without a placebo run-in, neither for drug-placebo nor for drug-drug differences (see Table 14.3). Although

Table 14.3 Effect Sizes of Studies With and Without Placebo Run-In

	No placebo run-in	Placebo run-in		
	r	r	z	$p(z)$
VERUM-PL	.18 (28)	.20 (62)	1.2	.23
TCA-SSRI	.01 (9)	−.01 (35)	−.50	.63

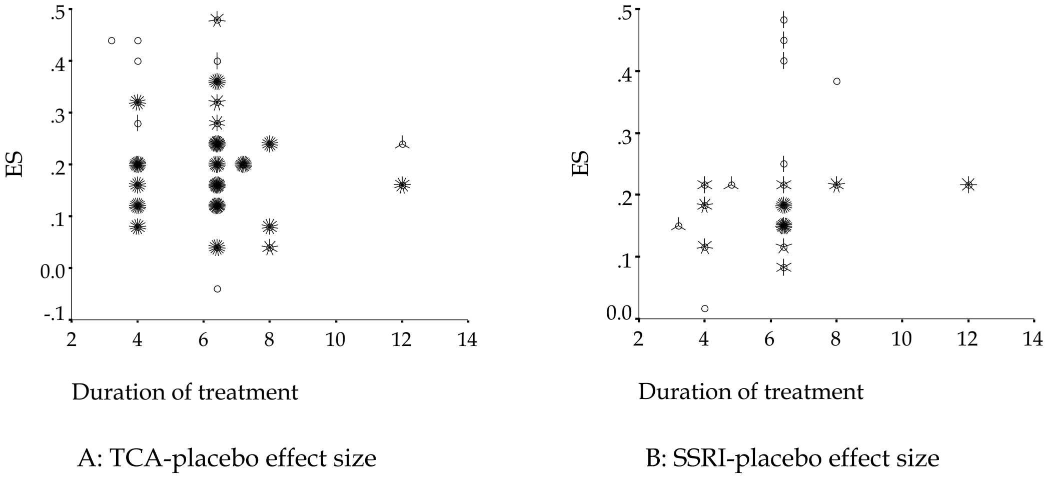
Note. For each r, the number of studies is given in brackets.

Table 14.4 Response Rates With and Without Placebo Run-In (Mean $\pm SD$)

	% Response Drug	% Response Placebo
Drug – Placebo		
No placebo run-in	52 ± 12	35 ± 10
Placebo run-in	48 ± 12	29 ± 11
	% Response Imipramine	% Response SSRI
Drug – Drug		
No placebo run-in	64 ± 9	63 ± 10
Placebo run-in	45 ± 15	46 ± 14

the number of responders was reduced among the studies with a placebo run-in, it was reduced in all treatment groups (see Table 14.4).

No correlation was found between effect size and *study duration* (see Figure 14.2, Panel A and B). The longer studies did not show higher effect sizes than the shorter ones. It must be noted, however, that there are only a few really short studies (< 4 weeks). Thus, we can neither conclude that longer studies show higher effect sizes, nor that there is no association.

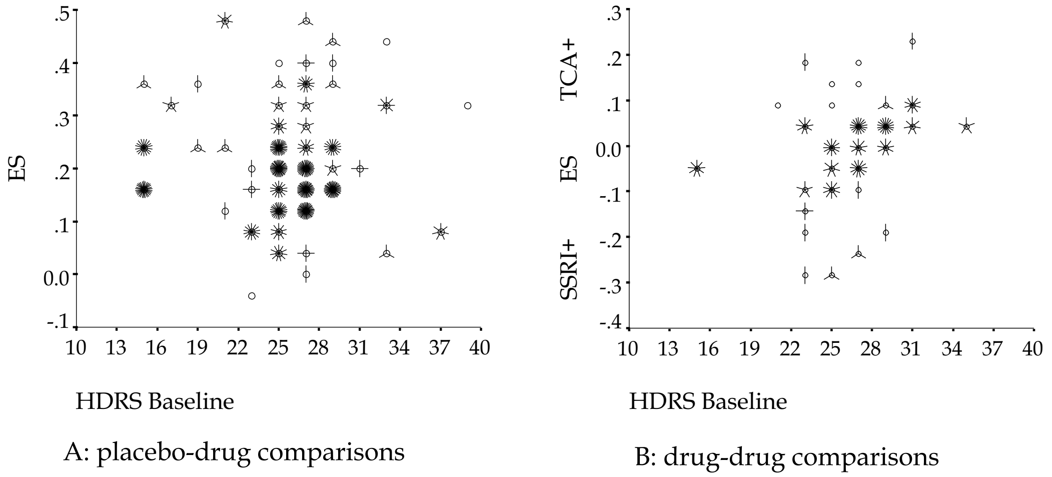


A: TCA-placebo effect size

B: SSRI-placebo effect size

Figure 14.2 Study duration (weighted by sample size) and effect size (ES).

Patients' *severity of depression* at inclusion was not correlated with effect size among the drug-placebo comparisons ($r = -.03$; n.s.; see Figure 14.3, Panel A). This finding is in contradiction to the widespread assumption, based on the concept of endogenous versus neurotic depression, that the superiority of drugs over placebo is more pronounced if patients show higher levels of symptoms (Feinberg, 1992). It may be the result of a missing sensitivity of a group mean score for baseline depression, used in meta-analysis; this would call for raw data analysis. On the other hand, there seem to be only few empirical proofs for a differential efficacy of antidepressants, except psychotic depression, as Montgomery and Lecrubier (1999) conclude in their review. If we consider drug-drug comparisons within our data, there is a low but significant correlation of .34. If a study included patients with a higher Hamilton baseline score, it was more probable to show a (minor) superiority of the TCA, whereas a lower baseline severity was rather associated with a better result of the SSRI (see Figure 14.3, Panel B).



A: placebo-drug comparisons

B: drug-drug comparisons

Figure 14.3 Sunflower plot of Hamilton Depression Rating Scale at inclusion (weighted by sample size) and effect size (ES).

These different results may be due to the presence or absence of a placebo cell. In fact, among the drug-drug-comparisons, there is a positive correlation of $r = .36$ between TCA-response and the Hamilton baseline score, whereas the SSRI response was not correlated with the Hamilton score ($r = .07$). Effect sizes of two cell studies (drug-drug comparisons) showed a correlation of .57 with the baseline Hamilton, compared to the zero correlation in drug-drug comparisons stemming from placebo-controlled studies ($r = .08$). In these studies, weak negative correlations with the Hamilton score can be found for all response rates (TCA $r = -.28$; SSRI $r = -.31$; Pl $r = -.29$). Thus, if a placebo is included in a study, the response in the single treatment cells is more likely to be negatively correlated with the severity of depression, if patients show lower levels of symptomatology, response occurs more often. We propose that this is the consequence of differences in drop-out handling. If there is no placebo cell in the protocol, it is likely that if symptomatology continues to be present, severe patients also stay longer on study medication because the treating physician knows that it is highly probable that the patients are on an active drug and not on a placebo. The shorter a patients adheres to the protocol, the shorter the time during which the drug has the opportunity to unfold its action.

The *response rate* under placebo is also a possible design variable, even if, logically seen, it belongs to the outcome variables of a study. Its correlation with effect size was $r = -.41$. The lower the placebo response, the higher the difference was between drug and placebo (see Figure 14.4).

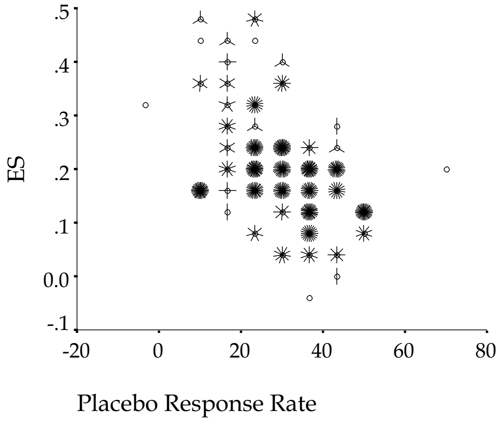


Figure 14.4 Sunflower plot: Placebo response (weighted by sample size) and effect size (ES) of drug-placebo comparisons.

14.7 DISCUSSION

This chapter focused on methodological aspects of controlled clinical trials on antidepressants. It presented the results of a publication-based meta-analysis of studies on acute therapy of depression with TCA (imipramine, amitriptyline) and/or SSRI (fluoxetine, fluvoxamine, paroxetine, sertraline) and/or placebo. Outcome was the effect size coefficient φ, which was based on intent-to-treat response rate differences.

Funnel plots indicate a publication bias, which is weak and probably does not affect the analysis of associations. We wish to consider our interpretations as hypothetical, as they are not based on a prospective design which aims at testing hypotheses and are subject to the problems of post hoc analysis by integration of different studies. Moreover, our interpretations base on data that were gained with a specific meta-analytical procedure and should be verified by other research strategies. Nevertheless, some statements can be made that may serve to better understand empirical results.

- Studies on imipramine yield higher effect sizes if they include only two treatment cells; this is in line with the results of Greenberg et al. (1992). We are inclined to attribute this to the earlier year of publication of the two cell studies and their less sophisticated research methodology; alternatively one can think of the status of an active or a control substance – or some other variables – being responsible for this difference in effect sizes.

- Smaller studies, or single center studies, have a higher probability of yielding or publishing higher effect sizes that decline if larger studies are performed.
- A placebo run-in with exclusion of placebo responders does not seem to have any effect on the outcome of a study and therefore becomes ethically questionable unless another argument is advanced for placebo run-in.
- In acute therapy of depression, there is no reliable association between the length of a study and its outcome.
- The correlation between severity of depression and response to treatment is linked to the design of a study: A placebo cell seems to decrease response in all treatment cells with increasing severity of illness. Presumably, if a placebo cell exists, patients drop out earlier, especially if symptomatology is more impairing.
- A negative correlation was found between placebo response and effect size. It is obvious that the correlation is not caused by a ceiling effect, as it is not only present in the margin values of the placebo response. For its interpretation, statistical or content aspects can be referred to. Statistically seen, the correlation between placebo response and effect size meets the expectation, as the difference between two sizes always correlates with both sizes at about .70. This so-called a (b-a) effect (van der Bijl, 1951) was discussed in psychophysiology in the context of the law of initial value (Myrtek & Foerster, 1986). Some authors (e.g., Curnow, 1987; Thompson, Smith, & Sharp, 1997) proposed methods to correct statistically for this problem, which is linked to the regression to the mean. One can, however, also find a non-statistical interpretation: There is one group of patients which responds well to placebo and another group who does not respond to placebo. The drug responders remain the same in both samples; this leads to varying differences in response rates, depending on the placebo response rate (Montgomery, 1999). Whatever the reason for this correlation, it should call into question the concept of additivity of the placebo- and drug-effect because this implies an independence of the response difference from the initial value.

Moreover, some hypotheses can be generated which could be prospectively investigated by empirical studies or raw data. It would be interesting, for example, to design a trial in which one part of the study is conducted within a placebo-controlled design while the other one consists only of a drug-drug comparison, if possible with documentation of the patients' and doctors' estimate of treatment allocation. Systematic differences of studies with and without a placebo control, which may be the consequence of differences in inclusion, medication and drop-out handling, could be investigated, as well as determinants and consequences of blindness. The comparability of placebo- and drug-controlled study designs is relevant for ethical reasons, since a placebo control is refused if a standard medication exists whose efficacy has been scientifically proven. This is only useful if a body of knowledge exists on the consequences of different study designs. The data presented here reveal the importance of this question, at least for antidepressant medication: If the SSRI had never been tested against placebo, and if statements on efficiency of the TCA had only been based on studies that investigated primarily TCA, the efficacy of the SSRI would be judged as higher, since older two cell studies on TCA yielded higher effect sizes.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (DSM–IV)* (4th ed.). Washington, DC.

Curnow, R. M. (1987). Correcting for regression in assessing the response to treatment in a selected population. *Statistics in Medicine*, *6*, 113–117.

Dickersin, K., & Berlin, J. A. (1992). Meta-analysis: State-of-the-science. *Epidemiologic Reviews*, *14*, 154–176.

FDA. (1978). Guidelines for clinical evaluation of psychotropic drugs – antidepressant and antianxiety drugs. *Psychopharmacological Bulletin*, *14*, 45–53.

Feinberg, M. (1992). Comment: Subtypes of depression and response to treatment. *Journal of Consulting and Clinical Psychology*, *60*, 670–674.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*, 71–79.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, *2*, 121–145.

Goodwin, F. K., & Jamison, K. R. (1990). *Manic-depressive illness.* New York: Oxford University Press.

Gotzsche, P. C. (1990). Bias in double-blind trials. *Danish Medical Bulletin*, *37*, 329–336.

Greenberg, R. P., Bornstein, R. F., Greenberg, M. D., & Fisher, S. (1992). A meta-analysis of antidepressant outcome under "blinder" conditions. *Journal of Consulting and Clinical Psychology*, *60*, 664–669.

Khan, A., Cohen, S., Dager, S., Avery, D. H., & Dunner, D. L. (1989). Onset of response in relation to outcome in depressed outpatients with placebo and imipramine. *Journal of Affective Disorders*, *17*, 33–38.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *2*, 719–748.

Möller, H. J., & Volz, H. P. (1996). Drug treatment of depression in the 1990s. An overview of achievements and future possibilities. *Drugs*, *52*, 625–638.

Montgomery, S. A. (1999). The failure of placebo-controlled studies. *European Neuropsychopharmacology*, *9*, 271–276.

Montgomery, S. A., & Lecrubier, Y. (1999). Is severe depression a separate indication? *European Neuropsychopharmacology*, *9*, 259–264.

Myrtek, M., & Foerster, F. (1986). The law of initial value: A rare exception. *Biological Psychology*, *22*, 227–237.

Petitti, D. B. (1994). Of babies and bathwather. *American Journal of Epidemiology*, *140*, 779–782.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

Thompson, S. G., Smith, T. C., & Sharp, S. J. (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine*, *16*, 2741–2758.

Trivedi, M., & Rush, H. (1994). Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology*, *11*, 33–43.

van der Bijl, W. (1951). Fünf Fehlerquellen in wissenschaftlicher statistischer Forschung [Five sources of error in scientific statistical research]. *Annalen der Meteorologie*, *4*, 183–212.

15

A Meta-Analysis of the Theory of Reasoned Action and the Theory of Planned Behavior: The Principle of Compatibility and Multidimensionality of Beliefs as Moderators

Ralf Schulze
Psychologisches Institut IV
Westfälische Wilhelms-Universität Münster

Werner W. Wittmann
Lehrstuhl Psychologie II
Universität Mannheim

Summary

Secondary analyses were conducted using 27 primary studies to assess the magnitude of relationships of intentions, attitudes, subjective norm, perceived behavioral control and their antecedents in the theory of reasoned action (Fishbein & Ajzen, 1975) and theory of planned behavior (Ajzen, 1985). As one of the purposes of these secondary analyses, the structure of belief components was explored for multidimensionality and the compatibility of the models' components was reliably assessed. The results were subsequently integrated under the random effects approach of meta-analysis. The magnitude of effects found in the theory of reasoned action fitted well within the context of hitherto published meta-analyses and showed strong overall relationships. Perceived behavioral control as a component of the theory of planned behavior was not found to be an important predictor of intentions in the present sample of studies, for which possible explanations are discussed. Moderator analyses of the compatibility of components resulted in consistent but somewhat low magnitude of effects. The dimensionality of belief components was of more importance for the relationships. Multidimensional representations have been shown to add approximately 10% of variance explained in attitude and subjective norm from belief based measures in comparison to traditional unidimensional measures. In contrast, the expectancy-value component could not contribute significantly to variance explanation of contiguous model components. The results are discussed in light of recent approaches in attitude structure and attitude–behavior research.

15.1 INTRODUCTION

Judgments of the utility of attitudes as a psychological concept have often been based on the relationship between attitudes and social behavior (Eagly & Chaiken, 1993, 1998). Whereas early approaches to an evaluation of the concept were quite optimistic (e.g., Allport, 1935), subsequent reviews questioned its utility as a predictor of overt human behavior and even suggested to abandon it as a scientific concept if consistency between attitudes and behavior could not be demonstrated (Wicker, 1969). This latter narrative review, in which it was concluded that there existed at most a slight relationship between attitude and behavior, has had a profound effect on the psychological research of attitudes. In the first half of the 1970s, attitudinal research was characterized by attempts to find explanations for the low correlations between attitudes and behavior reported in the review by Wicker (1969). Apart from methodological explanations, which we will address in the present study, new theoretical considerations have contributed to the question of when and how attitudes relate to overt behavior. One of the most important contributions of this type is the theory of reasoned action, introduced by Fishbein and Ajzen (1975).

15.1.1 The Theory of Reasoned Action and the Theory of Planned Behavior

The theory of reasoned action (TRA; Fishbein & Ajzen, 1975; Ajzen & Fishbein, 1980) connects attitude and its antecedents as well as its consequences in a systematic way and thus represents both a prediction and explanation model of overt volitional human behavior. In this model, depicted in Figure 15.1, variability in behavior is directly explained through intention, whereas the latter is predicted through attitude toward behavior and subjective norm. Simply put, the TRA stipulates that a person's behavior (B) is a direct (linear) function of her intention to act (I). As a consequence, overt behavior is considered as volitional in the TRA. Attitude toward behavior (A_B) exerts its assumed directive and dynamic influences mediated through intentions to act on behavior. Like-

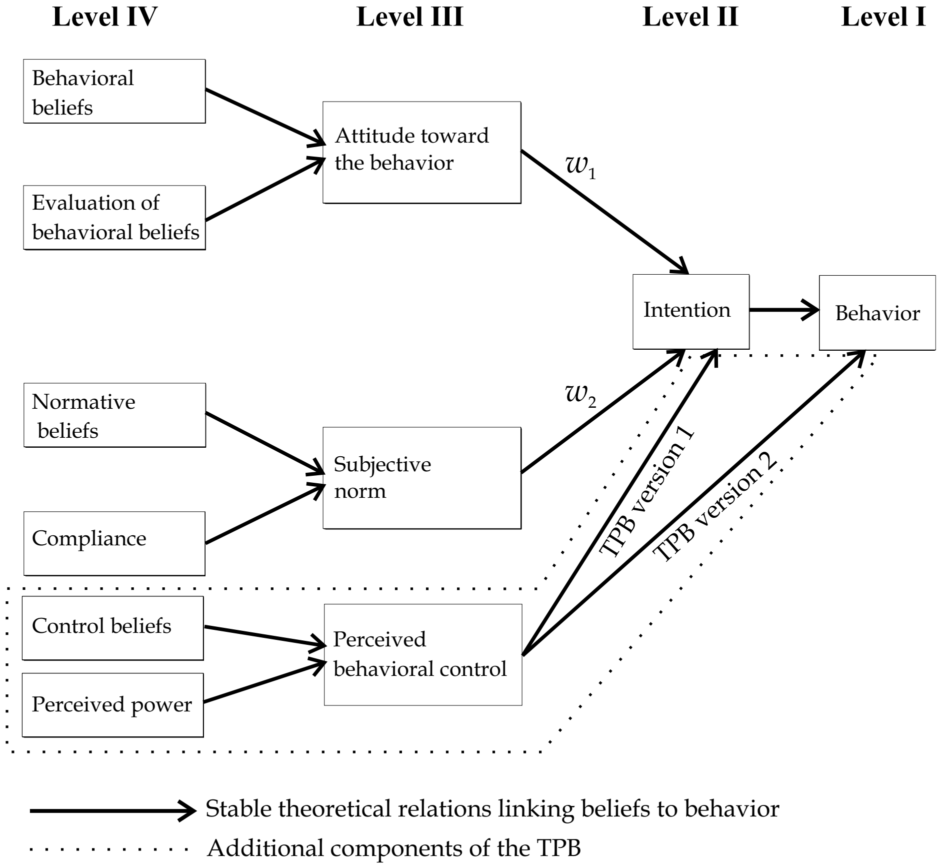

Level IV
Level III
Level II
Level I
Behavioral beliefs
Evaluation of behavioral beliefs
Attitude toward the behavior
w_1
Intention
Behavior
Normative beliefs
Compliance
Subjective norm
w_2
TPB version 1
TPB version 2
Control beliefs
Perceived power
Perceived behavioral control
Stable theoretical relations linking beliefs to behavior
Additional components of the TPB

Figure 15.1 The theory of reasoned action and the theory of planned behavior. Adapted from Ajzen and Fishbein (1980, p. 84).

wise, the impact of subjective norm (SN), which represents the influence of important others on a person's behavior is also mediated through intentions. The relationship between the components on levels I to III can be formally represented as (Fishbein & Ajzen, 1975, p. 301)

$$B \sim I = (A_B)w_1 + (SN)w_2,$$

where w_1 and w_2 represent the appropriate weights attached to A_B and SN, respectively. In practice, these relationships are ordinarily assessed by estimating the parameters of two OLS-regression equations separately. First, numerical indices of behavior are regressed on measurements of intentions, and second, a separate multiple regression of intention on attitude and subjective norm is performed to estimate w_1 and w_2 (for examples, see Ajzen & Fishbein, 1980; Fishbein, 1980). In the latter case, measures of overall predictive accuracy (R^2) are usually considered to judge the quality of the model.

On a fourth level, the TRA specifies the determinants of the level III components. Attitude toward behavior is conceived as a function of behavioral beliefs about consequences of the behavior in question and their evaluations, while subjective norm is a function of normative beliefs about important referents and the motivation to comply with these referents. By basing level III components on beliefs, according evaluations, and compliance, respectively, the TRA emerges as a model of *reasoned* action because behavior is ultimately founded in beliefs as cognitive aspects associated with behavior, its consequences, and important referents. There are several important points to note in the context of attitude formation and the formation of subjective norms. First, for a given attitude and person who holds this attitude, only *salient* behavioral beliefs are considered as determinants of attitudes, that is, beliefs that are accessible when an attitude object is encountered (Ajzen & Sexton, 1999). Due to limits of working memory capacity, a set of salient beliefs might be comprised of approximately five to nine beliefs on an individual level (Ajzen & Fishbein, 1980, p. 63). Pilot studies in which persons from the target population have to elicit their salient beliefs in free response format are common in applications of the TRA. The beliefs elicited in these pilot studies are often structured according to common sense criteria by the researcher and have ordinarily also to be reduced to a set of modal salient beliefs (Ajzen & Fishbein, 1980) that is intended to represent the set of beliefs salient in a given population. Secondly, for the prediction of level III components behavioral beliefs and evaluations as well as normative beliefs and compliance are thought to combine in a multiplicative form (Ajzen, 1996; Fishbein, 1963). This so-called expectancy-value model can be expressed as

$$A_B \propto \sum_{i=1}^{l} b_i e_i,$$

where the belief b_i represents the subjective probability that the attitude target is associated with a certain attribute. When the purpose of applying the TRA in a given situation is prediction of a specified behavior, the attitude target is an action and the attribute is a consequence of performing this behavior. The evaluation term e_i, in turn, can be thought of as the person's attitude toward the specified behavioral consequence. Thus, it is expressed on an evaluative continuum like good–bad. All products of beliefs and their evaluation are summed in the expectancy-value model to form a single composite which is used to predict the overall attitude in the TRA. This form of combining beliefs and evaluations is also applied to normative beliefs (nb_j) and compliance (co_j) as components of the TRA. A normative belief represents the subjective probability that a *specific* important referent thinks a person should perform a given behavior in question. The summed product of normative beliefs and motivation to comply with the behavioral prescriptions of specific others is used as a predictor of the according component on level III of the TRA:

$$SN \propto \sum_{j=1}^{m} nb_j co_j.$$

The TRA has been subject of debates and criticism over the past twenty years from various perspectives (for overviews, see Eagly & Chaiken, 1993; Jonas & Doll, 1996). Whereas one line of criticism addresses the conceptualization of components in the model, like intention as subjective probabilities of future behavior (e.g., Bagozzi & Kimmel, 1995), the main focus of most suggestions for improvements addressed the sufficiency of the model and thus centered around extending the set of variables in the TRA. One of these modifications is the Theory of Planned Behavior (TPB) (Ajzen, 1985), which attempts to extend the applicability of the TRA to behaviors in specific contexts, namely those in which behavior is not under volitional control (for a review, see Ajzen, 1991). In Figure 15.1, both the TRA and the TPB are depicted. The difference between these two theories lies in perceived behavioral control (PBC), control beliefs, and perceived power as additional components in comparison to the TRA. Perceived behavioral control is thought to reflect the perception of a person that a certain behavior is easy or difficult to perform. This perception might be routed in valid appraisal of external factors such as situational constraints that further or hinder the performance of behavior and may also concentrate on factors internal to a person, like necessary abilities or skills required for performance. Analogous to attitude and subjective norm, the antecedents of PBC are located on level IV of the model. Control beliefs are conceptualized as subjective probabilities about the presence of control factors that are of potential importance to perform a certain action or the strength of their association with a person. Again, these components are combined according to the expectancy-value model for the prediction of the level III component.

The influence of PBC on behavior is specified in two variants of the TPB (Ajzen & Madden, 1986). One path of influence is mediated through intentions to act and the other is directly headed to behavior. The influence on intentions reflects the tendency of persons to intend to engage in behaviors that are perceived to be under control over and above the directive effects of attitude and subjective norm. That is, persons intend to do things they perceive as easy to perform or past experience and anticipated obstacles, thought to be reflected in PBC, are in favor of performing the behavior (Ajzen, 1991). In another line of reasoning the direct influence of PBC on behavior in addition to a person's intention is interpreted as reflecting actual control over a behavior in question. The question of when and how these two paths of predicting behavior through PBC are theoretically and empirically supported is still a subject of attitude research (Ajzen, 1991; Eagly & Chaiken, 1993; Sutton, 1998).

15.1.2 Meta-Analyses of the TRA and the TPB

Both the TRA and TPB have received much attention and continue to stimulate most attitude–behavior research (Petty, Wegener, & Fabrigar, 1997). Table 15.1 provides an overview of mean effect sizes for various relationships reported in meta-analyses on these theories (see also Six & Eckes, 1996).

Although the comparison between these meta-analyses can only be qualitative because of partly overlapping study samples, the overall support for the relationships between model components reported in Table 15.1 is apparent and impressive. In addition to these meta-analyses, there are also traditional empirical reviews of the theories, some of which focus on specific research

Table 15.1 Meta-Analyses on the Theory of Reasoned Action and Theory of Planned Behavior With (Multiple) Correlations as Mean Effect Sizes

	Relationship under investigation			
Study	I – B	I – A	I – A + SN	A – B
Eckes and Six (1994)	.41 (96)	.42 (206)	—	.39 (396)
Farley, Lehman, and Ryan (1981)	—	—	.71 (37)	—
Godin and Kok (1996)[a]	.46 (26)	.46 (58)	—	—
Hausenblas, Carron, and Mack (1997)[b]	.47 (32)	.52 (23)	—	.39 (16)
Kim and Hunter (1993)	—	—	—	.47 (138)
Kraus (1995)	—	—	—	.38 (88)
Notani (1998)	.41 (45)	.45 (63)	—	.21 (19)
Randall and Wolff (1994)	.45 (98)	—	—	—
Ryan and Bonfield (1975)	.44 (35)	—	.60 (35)	—
Sheeran and Taylor (1999)[c]	—	.45 (32)	—	—
Sheppard, Hartwick, and Warshaw (1988)	.53 (87)	—	.66 (87)	—
van den Putte (1991)[d]	.62 (58)	.60 (88)	.68 (70)	—

Note. Number of studies in brackets. B = Behavior; I = Intention; A = Attitude; SN = Subjective norm.

[a]Only studies that focused on health-related behaviors were included. [b]Only studies that focused on exercise were included. [c]Only studies that focused on intentions to use condoms were included. [d]As cited in Eagly and Chaiken (1993).

fields like exercise research (Blue, 1995) or health behaviors (Conner & Sparks, 1996), and there are meta-analyses of the attitude–behavior relationship in related fields like advertising research as well (e.g., Brown & Stayman, 1992). All the reported and additional studies support the notion of very strong effects in the prediction of behavior and intention. Furthermore, reviews and meta-analyses on the behavior–PBC and intention–PBC relationships have also been published. For example, Godin and Kok (1996) report an average overall correlation between intention and PBC of .46 and a correlation with behavior of .39 on the basis of 58 and 26 studies, respectively. Despite this high overall correlation with behavior, only approximately 50% of the studies reviewed reported a significant incremental proportion of variance explained in behavior over and above the effect of intention. For the significant studies only, the mean incremental variance explained by PBC was 11.5%. The following studies also reported mean effect sizes on the relationships of PBC with intention and behavior, the number of studies reviewed is given in brackets: Hausenblas et al. (1997) behavior–PBC = .45 (8), intention–PBC = .43 (10); Notani (1998) behavior–PBC = .24 (45), intention–PBC = .31 (63); Sheeran and Taylor (1999) intention–PBC = .35 (24).

In sum, the TRA and TPB have received strong overall empirical support for important relationships of model components. Although it has been repeatedly shown that mean effect sizes are strong for the various relationships it must be added that most meta-analyses also reported heterogeneous effects. Moderator analyses have therefore also been performed to test for the moderating effect of miscellaneous variables. Some of these potential moderators stemmed from psychological reasoning, like attitude accessibility, strength or certainty (e.g., Kraus, 1995), but there have also been methodological considerations to explain correlational differences between studies not only in the attitude–behavior relationship but also in the relationships of the components of the models in general. We will turn to two specific moderators that are addressed in the present study after extensions of the theories have been outlined in the following section.

15.1.3 Extensions of the TRA and the TPB

The sufficiency of the TPB and the TRA has been repeatedly questioned and additional important variables have been proposed, at least in specific contexts (for a review, see Conner & Armitage, 1998). Moral norm and self-identity seem to play a major role here, as evidenced by their inclusion in the attitude–behavior composite model of Eagly and Chaiken (1993, 1998) or the theoretical framework of Triandis (1980), for example.

Moral norm on the one hand is concerned with the perception of a person that a certain behavior or its consequences are inherently wrong or right apart from judging it with respect to personal utility (behavioral beliefs) or social influences (normative beliefs). A person may thus feel a moral obligation to perform a behavior according to internalized moral standards. Consequently, this component has been added to the TRA by several researchers and was found to add to the prediction of intention in addition to attitude and subjective norm in most applications (for a review, see Manstead, 2000). In contrast to this relatively consistent research evidence, there is considerable heterogeneity as far as the location of moral norms in the TRA or TPB is concerned. Whereas initially personal norms were conceptualized as a second dimension of normative influences on behavior (Fishbein, 1967) on level IV of the TRA and therefore not necessarily qualitatively different from social norms, they are usually introduced in applications as a component on level III in addition to attitude and subjective norm (e.g., Gorsuch & Ortberg, 1983). This “shift” resulted from a re-conceptualization of personal norms that is more sharply focused on the moral implications of a behavior in question. Accordingly, these personal moral and ethical standards were more precisely termed moral norms. Despite this focus on moral aspects, it is neither theoretically nor empirically clear whether this component should be regarded as an antecedent variable of attitude or as an addition to it (cf. Sparks, Shepherd, & Frewer, 1995; Parker, Manstead, & Stradling, 1995).

Self-identity is another component that has been added to the TRA and TPB to enhance their explanatory and predictive power of behavior. The following two examples illustrate the meaning and varying emphasis placed in definitions of this concept: According to Sparks (2000, p. 35), self-identity is defined as a person's self-concept, that is, relatively enduring characteristics that a person ascribes to herself, whereas Conner and Armitage (1998, p. 1444) define self-identity as the "salient part of an actor's self which relates to particular behavior". Furthermore, Sparks (2000) pointed to the fact that expressions of self-identity may also incorporate moral norms. The two components added to the TRA and TPB are therefore not clearly distinct. Despite these conceptual difficulties self-identity has a relatively fixed hypothesized position in the TRA and TPB. It is mostly assumed to be associated with attitude in the TRA and TPB, and has also shown to influence intention in addition to attitude and subjective norm (for reviews, see Conner & Armitage, 1998; Eagly & Chaiken, 1993; Sparks, 2000).

15.1.4 Multidimensionality of Beliefs

In the context of the expectancy-value model the summation of beliefs and according evaluations, for example, includes all modal salient beliefs determined in a pilot study and involves no weighting of these parts of the composite. This amounts to a highly restrictive unidimensional model with equal component loadings of all parts to be summed that has only occasionally been explicitly tested in applications of the TRA and TPB. The potential failure to map a multidimensional belief structure in appropriate components may cause serious consequences for the relationship between level IV and level III components, which have been judged as relatively low and "somewhat disappointing" (Ajzen, 1991, p. 192).

In fact, alternatives to an unidimensional representation have been repeatedly proposed, even in an early statement of the TRA (Fishbein, 1967). Another early approach can be seen in the work of Scott (1969) who introduced measures of structural properties of cognitions, one of which was dimensionality, thought to represent "the space utilized by the attributes with which a person comprehends the domain" (Scott, 1969, p. 263).

Perceived behavioral control is one example of a potentially multidimensional component. There is considerable theoretical debate about the subdivision of PBC into self-efficacy and controllability. There is also empirical evidence that these subcomponents can be successfully represented in a two-dimensional model as well as that they differentially predict intention to act and behavior (Armitage & Conner, 1999; Conner & Armitage, 1998; Sparks, Guthrie, & Shepherd, 1997; Terry & O'Leary, 1995). Furthermore, normative beliefs and compliance may also be of a multidimensional structure. It seems quite possible, for example, that normative beliefs associated with family members as opposed to normative beliefs associated with friends form different components and these two belief sets might contribute independently and differentially to the prediction of other model components. Burnkrant and Page (1988) have in fact shown that a two-factor structure with more closely related referents like friends and spouses loading on one factor and parents as well as employers on the other hand loading on another factor shows a significant improvement in fit of the model and an improved prediction of contiguous model components (see also Grube, Morgan, & McGree, 1986). Most research on multidimensionality of beliefs has been conducted on the dimensionality of behavioral beliefs as the basis of attitudes (Bagozzi, 1981a, 1981b; Grube et al., 1986). From a theoretical viewpoint, Schlegel and DiTecco (1982) argue that multidimensionality may be more prevalent in domains that can easily be described by many characteristics and where persons under investigation have a differentiated knowledge structure. As a consequence, single representations like the expectancy-value model are supposedly not capable to map such a differentiated structure in a single score. Indeed, for a study on marijuana use they provided evidence that the dimensionality increased with more experience and presumably more knowledge about marijuana in different user groups. In another large study on non-medical drug use they replicated this finding and, more important, they showed that the multiple correlations of unidimensional representations with behavioral intention and behavior were lower as for the multidimensional case. This effect was even more pronounced for users with more experience, that is, those with a more complex representation of attitude (Schlegel & DiTecco, 1982).

In sum, there are theoretical reasons as well as empirical evidence that the exploration and testing of multidimensionality of the components of the TRA and TPB is a promising route to better understand and predict level III components and ultimately behavior. Normative and behavioral beliefs can be regarded to represent multiple (two) dimensions or domains of beliefs themselves that are separated for theoretical and practical reasons. As has been shown, these dimensions may also be composed of a set of subdimensions. Normative beliefs can be partitioned into groups of persons that differentially predict the overall perception of normative influences on intentions. Behavioral beliefs can also be subdivided into context specific belief sets or more general groups of beliefs that map different facets of utilities. Whereas utility frequently is associated with more instrumental or material outcomes of behavior, from the multidimensional perspective it might encompass several dimensions from solely material outcomes to outcomes of ideational value, moral relevance or whatever dimension of worth is prevalent in a given context. Especially in domains in which differentiated knowledge is prevalent for a given sample, multiple differentiable dimensions are expected to emerge. It is suggested that these different dimensions may all be represented and employed to predict behavior and its antecedents.

The potential benefits of a multidimensional approach are therefore manifold. First, it is explicitly tested or explored whether one dimension is sufficient to represent the belief structure in a given context and multiple dimensions are regarded as alternative representations if a unidimensional measurement model fails to fit. Second, in the case of more than one dimension the results can reveal *what* distinguishable dimensions of worth are relevant in a given context, where the contrast between instrumental and moral beliefs is only one possibility. Third, with multiple dimensions it is possible that persons actually have an inconsistent belief basis that would go unrecognized in a composite score. A multidimensional approach at least offers the chance to uncover such inconsistencies. Fourth, if multiple dimensions are given, then the prediction of attitude or intention may be enhanced by this approach or the failure to successfully predict intention and behavior through level III components may be explained by structural properties of the belief basis.

15.1.5 The Principle of Compatibility

Although one of the results in the influential review by Wicker (1969) was the often cited low correlation between attitude and behavior, it should be noted that he also presented some explanatory factors that were hypothesized to influence this relationship. With reference to the work of Fishbein, he introduced the specificity of attitudes as one of these factors, which is one aspect of the *principle of compatibility*[1]. He argued that for different levels of specificity of attitudes and behavior, only low correlations are expected whereas with equal specificity he anticipated high correlations. The prototypical case of a specificity mismatch is seen in a measure of global attitudes and a specific behavior. For example, the low attitude–behavior relationships in the often cited study of LaPiere (1934) was ascribed to such a mismatch of levels of specificity (Stroebe, Eagly, & Ajzen, 1996).

Fishbein and Ajzen (1974, 1975) elaborated on this moderator and presented a systematic approach to construct more general measures of behavior which they termed multiple act criteria in contrast to the more specific single act criteria of behaviors. Multiple acts are, in essence, aggregates of single acts that consist of specific behaviors, performed in various contexts and points in time. Furthermore, they did not only specify principles for the construction of multiple act criteria but also stated different compatibility characteristics with respect to which components of their models could match or mismatch. More specifically, Fishbein and Ajzen (1975) distinguished four dimensions: Target, action, context and time (TACT), where components have to match as a prerequisite for strong relationships. The target in this classification system is the object at which a behavior is directed, action is the behavior itself, context and time are the environment in which the behavior takes place and the point in time when the behavior is shown, respectively. Although most treatments of the principle of compatibility focus on the attitude–behavior relationship, all the components from level I to IV of the TRA and TPB can be characterized by

[1] Originally (Fishbein & Ajzen, 1975) termed "principle of correspondence". In accordance with Ajzen (1988) and Eagly and Chaiken (1993) the term "principle of compatibility" is used here.

the TACT-dimensions. Since level II to IV components are of utmost importance for the present study, we focus on these components in the following.

To illustrate the principle of compatibility more concisely, imagine an attitude toward eating low-fat food in the next two weeks measured with the semantic differential technique. Here, the action is eating which is targeted towards low-fat food. Whilst eating is a rather specific action, low-fat food is a category of food that includes a great deal of products like skimmed milk, salads, fruits and the like. The situations and circumstances under which eating takes place are not specified, so the context component is regarded as general whereas the time component is restricted. Here, the attitude should be related to eating taking place within the next two weeks. Now imagine a set of beliefs that is intended to be compatible to this attitude. Ideally, this set should be restricted to beliefs that address the personal consequences of eating low-fat food in the next two weeks. The time period should therefore be specified in exactly the same way as in the attitude measure, just as the action component. For the more general attitude components of target and context, there are at least three variants to specify these in the formulation of belief items. First, it is possible not to formulate anything about these components, so that they are as general as in the attitude measurement. Second, there is the possibility to specify many specific exemplars in the formulation of belief items as long as the set of items encompasses all conceivable contexts in which the behavior can be performed or all targets a behavior is directed to. Finally, prototypical contexts and targets may be chosen. Although the latter two options may in principle be realized, it is obvious that in individual cases it is very difficult to decide whether a given set of belief items is indeed prototypical or general enough to be compatible with an unspecified TACT-aspect in another component of the model. As a consequence, it is argued that the question of match or mismatch of components is actually a *matter of degree* and not a matter of kind. It should furthermore be noted that as a consequence, the assessment of compatibility is not at all an easy or trivial task. For a valid assessment of the compatibility of the TACT-dimensions it seems necessary to consider *all* items used in a study and assess their level of specificity.

From the first extensive formulation of the TRA in 1975 on, Fishbein and Ajzen advocated the principle of compatibility as one of the most important moderators of the relationships between model components and especially the attitude–behavior relationship. Not only does its fundamental idea have implications for attitude research, but it is also relevant for research in the psychology of personality (Ajzen, 1988; Sherman & Fazio, 1983). Moreover, the principle can also be taken as a methodological tool for successful validation and modeling strategies in general (Kirkpatrick, 1997; Nesselroade & McArdle, 1997; Wittmann, 1988).

Notwithstanding its general nature and many successful applications, the principle of compatibility has occasionally been regarded as a merely methodological tool and it has been stated that “it is not very exciting from a psychological point of view” (Millar & Tesser, 1992, p. 278), and that “it was formulated without much attention to the underlying psychological mechanisms”

(Ajzen & Sexton, 1999, p. 130). There are, however, elaborated accounts of the principle's theoretical underpinnings. A first approach can be seen in the work of Millar and Tesser (1986, 1992). They proposed the mismatch-model in which it is stated that the prediction of behavior from attitudes will be poor if the focus on affect versus cognition in attitude formation and during performance of behavior is different. Accordingly, they proposed and empirically demonstrated that high relationships between these components can be observed under matching conditions. In a similar vein, Ajzen (1996) introduced the notion of belief equivalence during the expression of attitude and behavior, which was extended to the so-called principle of belief congruence (Ajzen & Sexton, 1999). But perhaps the most elaborated approach was recently presented by attitude representation theory (Lord & Lepper, 1999), which shows remarkable similarities to the principle of compatibility. In sum, all these theories demonstrate the substantial psychological basis of the principle.

Empirical tests of the principle of compatibility have mainly focused on general attitudes and their failure to predict single act criteria (for examples, see Ajzen & Fishbein, 1977; Jaccard, King, & Pomazal, 1977). Evidence from two meta-analytical studies (Kim & Hunter, 1993; Kraus, 1995) suggests that compatibility of attitude and behavior is an important moderator of the attitude-behavior relationship. However, there are a few shortcomings with these meta-analyses. For example, it is not clear how compatibility of measures was assessed in the Kim and Hunter meta-analysis. Usually, only few examples from the questionnaires or interviews used in the original studies are reported. The categorization into low, moderate and high match groups as done in the Kim and Hunter meta-analysis can only be based on the examples reported. In the face of the difficulties in assessing compatibility of components outlined above, this can be regarded as a very crude measure. Results reported in Kraus' meta-analysis were based only on a very small subset of studies (8 out of 88), which directly investigated the effect of compatibility of attitudes and behaviors. In sum, though consistent empirical support was presented for the principle of compatibility, a stringent meta-analytical test of the hypothesized moderating effect of compatibility, based on a reliable measure that maps the various TACT-dimensions in one or several scores, is not yet available.

15.1.6 Aims of the Study

In the present study, we pursue several objectives. First, we will evaluate the TRA and TPB through the use of meta-analysis after performing secondary analyses of original data. The relationships between several model components will be assessed and compared with respect to published meta-analyses (see Section 15.1.2). The results will give some indication whether the effects of unpublished studies do actually differ in comparison to the results of published studies as assumed in the file-drawer hypothesis (Rosenthal, 1979, 1991). Second, we will assess the potential multidimensionality of components on level IV of the models and PBC. In addition, we will give some indication on what dimensions emerged in the assessment of multidimensionality and assess their predictive power for other components in comparison to unidimensional representations. Third, compatibility of the components will be employed as a predictor to explain the variability of effect size variances under the random effects model of meta-analysis. In contrast to other meta-analytical tests of this moderator, we will not only focus on the attitude–behavior correspondence and compatibility, but test whether the potential moderating effect also extends to the relationship of other components of the TRA and TPB, an effect we expect from the generality of the principle.

15.2 METHOD

The present study represents a mixture of secondary analyses and a meta-analysis. In a first step, secondary analyses were performed on all available data sets in order to check the quality of the data. For example, we explored the distributional properties of the variables, and computed the relevant statistics for the subsequent meta-analytical step. Every step of the secondary analyses, presented in more detail in Section 15.2.2, applies to every single study whereas the following meta-analytical steps serve to integrate the results.

15.2.1 Selection of Studies

All analyzed data sets pertain to heretofore unpublished studies submitted as diploma theses at a German University and had to meet the following criteria:

1 A complete report of the study, including all measurement instruments, had to be available.

2 Raw data of all studies had to be available in order to perform all the steps of the secondary analyses.

3 The TRA or the TPB had to serve as theoretical background for the studies.

The pool of studies was not systematically sampled from a population of unpublished studies. Generalizations to unpublished studies on the Fishbein and Ajzen models are therefore not warranted, albeit the results will at least shed some light on the effects to be expected of so-called file-drawer studies.

The final sample included 27 studies with a total number of 4499 respondents. Selected study characteristics are reported in Table 15.2. The overall mean age for respondents is 24.8 and the mean number of respondents per study is 166.6. Fifteen of the studies investigated PBC as an additional component and were therefore classified as implementing the TPB. In all studies the semantic differential technique was used as measurement instrument to directly assess the respondents' attitudes. In addition, item forms and wordings for the other model components were used as recommended by Ajzen and Fishbein (1980) for the TRA and by Ajzen (1991) as well as Ajzen and Madden (1986) for the TPB. The adherence to the recommendation for the TPB led to a

Table 15.2 Selected Study Characteristics

Study	N	Mean age	Attitude topic	Attitude toward	Theory
1	157	30.2	Taking on a higher position in a company	Act	TPB
2	176	23.5	Moving to East Germany after passing the exam	Act	TRA
3	298	27.0	Participating on a training course in a company	Act	TPB
4	210	22.4	Having an abortion	Act	TPB
5	112	24.4	Pursuing a career after having a baby	Act	TPB
6	180	43.5	Becoming a teacher	Act	TPB
7	98	35.9	Eating health food	Act	TPB
8	232	27.1	Specific German company from heavy industry	Object	TRA
9	300	34.2	Credit cards	Object	TRA
10	157	25.6	Assessment Center	Object	TPB
11	110	25.6	Having vocational education after the exam	Act	TRA
12	88	24.1	Making a decision concerning the statutory basis of the German Reunion	Act	TRA
13	212	24.5	Right of asylum	Object	TPB
14	144	14.6	Doing "something against" foreigners	Act	TPB
15	121	25.3	Jobs in East Germany	Object	TRA
16	343	15.7	Various disciplines taught in school	Object	TRA
17	111	26.3	Participating on a training course in a company	Act	TPB
18	191	25.0	The study at university with respect to practical applications	Object	TPB
19	85	18.4	Going to a vocational school	Act	TRA
20	112	20.6	Working with the computer	Act	TRA
21	106	38.2	Paying with credit cards	Act	TPB
22	42	17.1	Work experiences	Object	TRA
23	269	19.3	Serving in the army	Act	TRA
24	114	21.5	Deciding to become a career women vs. housewife	Act	TRA
25	104	20.7	Participating on a demonstration	Act	TPB
26	167	18.5	Studying at university	Act	TPB
27	260	21.1	Studying at university	Act	TPB

Note. N = total sample size, TRA = Theory of Reasoned Action, TPB = Theory of Planned Behavior.

mixture of controllability and self-efficacy items in 9 of the 15 studies that employed the TPB (see Section 15.1.4). The remaining 6 studies employed only fewer than 3 items to assess perceived behavioral control, all of which were controllability items. In every study a specific set of items was constructed, first pretested in a pilot study for applicability to a larger pool of subjects from the same population. Modal salient beliefs were also determined in these pilot studies to assure relevance of the belief items for the respective sample of respondents. As only two studies reported results on the relationship of the model components to overt behavior, this aspect of the models will be left out in the following sections. This also applies to control beliefs and perceived power, which were assessed in only two studies.

15.2.2 Secondary Analyses

The first step was to adjust the data from the studies under investigation to the recommendations proposed by Ajzen and Fishbein (1980). This included rescaling of items and computation of expectancy-value components, if necessary. Since the issue of unipolar versus bipolar scaling is still under debate (Eagly & Chaiken, 1993; Sparks, Hedderly, & Shepherd, 1991), items were scored as proposed by Ajzen and Fishbein (1980) to provide a fair test of the theories.

In order to keep the number of variables to analyze in subsequent steps at a reasonable level and to assess potential multidimensionality of the components, items were compressed via principle component analyses with one and multiple factor solutions, if indicated. All components of the TRA and TPB were subjected to this procedure, apart from components which were assessed with fewer than four items. Component scores were thereby calculated for all subjects for further computations. The one-factor solutions correspond to unweighted sum variables of multiple item scales usually employed in analyses of TRA and TPB applications, but are superior in the sense that they preserve a maximum of variance of the items to be aggregated. In addition to the one-factor principle component analyses, multiple factor solutions were explored and implemented in cases where conventional statistical criteria like the eigenvalue-greater-than-one-rule and the scree-plot indicated that more than one component could be extracted. Moreover, attention was also paid to the psychological significance of the solutions. All multiple component solutions were rotated after extraction with varimax rotation to achieve simple structure of the loading matrices. One exception from the outlined procedure was the case of attitude measurement with the semantic differential technique, which was employed in all studies. Here, multiple components were always extracted to obtain scores for only the evaluative dimension that represents attitudes (Fishbein & Ajzen, 1975). In all studies factored separately, this evaluative dimension clearly emerged after rotation of the components.

As reliability estimates for the components we used a formula based on the eigenvalues given by Cliff (1988) which he also criticized for its strong assumptions. Since no reliability estimates of the single items were available, we were not in a position to perform reliable component analyses for better estimates of reliability (see Cliff, 1988; Cliff & Caruso, 1998). The mean reliabilities we computed were acceptable and well above .75 for all components except PBC. For the studies under review, PBC showed a mean reliability estimate of .63, which, though not unacceptable, is well below the reliabilities for the other components.

The results of multiple component analysis revealed several results worth mentioning. First, intentions and subjective norm, as operationalized according to the recommendations of Ajzen and Fishbein (1980), consistently showed only one component in all studies. This was mainly due to a focus on specific behaviors in the various studies in the case of intentions, and mostly few or only one item to measure subjective norm. In contrast, 9 of 15 TPB-studies employed a mixture of controllability and self-efficacy items which resulted consonantly in two components for all these studies. This result stands in agreement with similar attempts to separate these two components (Conner & Armitage, 1998; see Section 15.1.4). Second, multiple component solutions of behavioral beliefs and according evaluation of behavioral beliefs showed remarkable similarities in structure which also mirrored the structure of multiple component analyses of the according expectancy-value product terms that were factored separately from the former. Despite the fact that some evaluation of behavioral belief items loaded highly on one component and the according behavioral belief items did not load as equally high on the respective component in a separate analysis of behavioral beliefs and vice versa, this did not vitiate the similarity of structure as far as the interpretation of the components is concerned. The structure of beliefs that emerged was partly specific for the behavioral domains addressed in the studies, like several stress and strain effects of participating in an assessment center (study 10) or various specific health consequences of consuming health food (study 7), for example. On the other hand, there were also noteworthy similarities of interpretation of factors *across* studies. These similarities pertain to principle component analyses of behavioral beliefs that lead to partitioning of beliefs in economic/material, moral, and self-related beliefs in most of the studies. The economic components consisted of mainly utilitarian beliefs in the sense of monetary consequences of certain behaviors like earning or saving more money when moving and working to West or East Germany (e.g., studies 2 and 15), for example. Another facet was found in the more ideational or moral aspects of the utility of behavioral consequences by the participants. Here, beliefs can be exemplified by the violations of ethical rules through discrimination of ethnic minorities (e.g., studies 13, 14 or 25) or burdening of future generations through environmental pollution (e.g., study 8). The last facet of self-related beliefs is comprised of beliefs that deal with self-realization or self-esteem, that is, beliefs about behavioral consequences that touch upon a person's needs, interests, or self-esteem. This latter component did emerge in all studies with more than two components and most concisely in studies on behaviors in a learning environment like universities or training departments of a company (e.g., studies 1, 3, 26, 27), but differs somewhat in meaning from the notion of self-identity outlined in Section 15.1.1. The components of this facet extracted from the studies in the secondary analyses focused in meaning more on outcomes that enhance or undermine self-esteem (e.g., the feeling of pride as a behavioral consequence) and to a far lesser extend on issues of personal or social identity. In sum, subjective probabilities and evaluations of behavioral consequences as well as expectancy-value components showed a similar component structure supplemented by domain specific components that differed between studies. Remarkably, the independent components found across studies resemble components that have been added to the TRA and TPB (see Section 15.1.1) to enhance explanation and prediction of behaviors in certain domains.

The next step of the secondary analyses was to compute the linear relationships of the components using multiple regression, where R^2 was recorded as effect size. Since R^2 is a biased estimate of the coefficient of determination in the population and standard errors were needed for subsequent steps, a bootstrap procedure was applied to compute a bias-corrected R^2 and according standard errors using 300 bootstrap resamples in each study (for details of this procedure, see Efron & Tibshirani, 1993). Furthermore, in order to detect violations of the model assumptions hierarchical regressions were computed. Following the recommendations of Evans (1991), expectancy-value components were added to behavioral beliefs and their evaluations. Incremental variance explained by these components was recorded and tested for significance with hierarchical F tests. To test the significance of explained variance through additional components, resulting significance levels from the hierarchical F tests were integrated as described by Rosenthal (1991).

15.2.3 Assessment of Compatibility

To assess the compatibility of the models' components, three undergraduate psychology students rated all items of contiguous components on the compatibility dimensions on a five point scale in the last step of the secondary analyses. Students were trained beforehand to become acquainted with the TACT-dimensions. The training consisted in thorough reading and discussion of an extensive manual on the theoretical background of the principle of compatibility. The manual summarized the relevant literature on this topic (e.g., Ajzen & Fishbein, 1977, 1980; Fishbein & Ajzen, 1975) and explained the principle with prototypical examples of items, which resembled but were not identical to the items of the studies under investigation. Apart from the more theoretically oriented part of the training a manual of rules was also prepared which explained how response option of the ratings should be used by the raters. All rules were explicitly stated, illustrated with concrete examples, and verbally explained. This manual of rules was subdivided into the following four parts, which matched the tasks to be fulfilled to rate the compatibility of the components:

1 Rules for ratings of the specificity of a single component on the TACT-dimensions.

2 Rules for ratings of the compatibility of two model components.

3 Rules for ratings of the joint specificity of two model components.

4 Rules for ratings of the compatibility of two combined model components.

As can be seen by the structure of the four parts, raters had to rate the specificity of the items first. This step was introduced to force focus on the specificity of the model components on every TACT-dimension separately before compatibility ratings were conducted. This step was of special importance for components that were assessed with several items in all studies, like behavioral beliefs for example. In these cases, the specificity rating of the target, for example, applied to the whole group of behavioral belief items. The ratings of the second step were only conducted after the first step was applied to both components to be rated. Steps 3 and 4 were only applicable to level III and IV components and served to structure ratings of the compatibility of two components of level IV, behavioral beliefs and according evaluations or normative beliefs and compliance, and one component on level III, namely attitude and subjective norm, respectively.

In addition, several aspects of the rules of compatibility ratings are worth mentioning. First, the ratings for the components which were assessed with a set of items and were therefore subjected to principal component analysis focused only on items loading higher than .30 on the respective component. This rule was introduced to prevent ratings to be influenced by items that do not substantially contribute to the components scores to be used in the regression analyses. Second, in cases where, for example, the context of action was not specified when measuring attitude with the semantic differential technique and many behavioral belief items were specific with respect to the context of action, there was sometimes disagreement between raters. The cause for these disagreements was the difficult decision task for raters to judge whether the ensemble of specified contexts in behavioral belief items was broad enough to be compatible to an unspecified attitude. No objectively determinable criteria were available to resolve such disagreements, so a final discussion session was held with all raters to focus on and discuss such disagreements. Finally, it is important to note that the raters were, at the time of rating the questionnaires, not knowledgeable of any result of the studies, to prevent ratings to be influenced by such knowledge.

The degree of agreement of the raters in the final ratings was assessed as intraclass reliability coefficients with raters as fixed and studies as random factors (Shrout & Fleiss, 1979). The reliability estimates for overall average ratings of compatibility for the three raters are presented in Table 15.3. With few exceptions, all intraclass coefficients for the specificity ratings not reported in Table 15.3 were at least .65, with more than 60% of these coefficients above .80. The exceptions were ratings for context and time specificity of evaluation of behavioral beliefs and compliance, context specificity of behavioral beliefs, subjective norm, and normative beliefs, as well as time specificity of perceived behavioral control. For all these components zero reliability estimates resulted from missing variance in the ratings, which actually indicates *perfect* agree-

Table 15.3 Reliabilities of Overall Compatibility Ratings

Relationship	Reliability estimate
Intention – Attitude	.86
Intention – Subjective Norm	.73
Intention – PBC	.88
Attitude – BB	.70
Attitude – EBB	.90
Attitude – BB + EBB	.79
Subjective Norm – NB	.23
Subjective Norm – CO	.93
Subjective Norm – NB + CO	.67

Note. The number of studies is given in brackets. Reliabilities were computed as intraclass coefficients on the basis of the ratings from three raters. PBC = Perceived behavioral control; BB = Behavioral beliefs; EBB = Evaluation of behavioral beliefs; NB = Normative beliefs; CO = Compliance.

ment between the raters. As a consequence, this missing variance will also lead to an exclusion of these ratings from the moderator analyses.

As can be seen in Table 15.3, reliabilities were acceptable with the exception of the compatibility ratings between subjective norm and normative beliefs. This result can be traced back to a highly restricted range of compatibility ratings for these components. Although intraclass coefficients were well above .80 for the specificity ratings of subjective norm and normative beliefs, the compatibility between these components was essentially rated as nearly perfect for all studies. On the five-point scale from 1 (no compatibility) to 5 (perfect compatibility) more than 50% of the studies showed scores of 5 and the remaining studies had mean scores equal to or above 4. As a result, the compatibility of these components could not be employed as a moderator in subsequent analyses.

15.2.4 Meta-Analytical Procedures

In all previous meta-analyses concerning the TRA and TPB relationships between the components of the models were assessed by the Pearson product-moment coefficient r. Reported multiple correlations in the original studies to be synthesized have usually been treated as if they were r. From a statistical viewpoint, this is inappropriate since these statistics have different sampling distributions and standard errors. In order to use a common effect size estimate, the coefficient of determination R^2 was chosen in the present study. Although R^2 and similar measures of variance explained have been criticized as effect size estimates because these measures do not indicate the sign of an effect (Hedges & Olkin, 1985), this criticism does not apply in the context of the TRA and TPB, as long as linear prediction is not accomplished through counterintuitive effects. If, for example, a favorable attitude towards having an abortion were negatively related to the intention of actually having an abortion, R^2 would be misleading as an indicator of the effect. Special care was given to detect such counterintuitive effects, but none were encountered in any of the secondary analyses. Another problem with measures of variance explained like R^2 lies in the estimation of its standard error, which plays an important role in meta-analysis as a component of the weights for the studies. In the present study we have used the bootstrap estimates of the standard error for the R^2s that were computed in the secondary analyses.

Another decision to be made in the present meta-analysis pertains to the assumption of a fixed versus random effects model. The distinction between these models is an important one for meta-analytical methods, as evidenced in several chapters of this book. In the fixed effects approach it is hypothesized that all studies under investigation estimate a common effect size but in the random effects model true differences in effect sizes between studies are assumed (Hedges, 1983; Hedges & Vevea, 1998). As a consequence, the observed variance in estimates of effect size parameters is attributed to errors of estimation in the fixed effects model, whereas in the random effects model the observed variance of effect sizes is partitioned into variance due to true differences in effect sizes on one hand and variance due to errors of estimation on the other. Strong arguments have been put forward in the recent literature on meta-analysis in favor of the random effects model (e.g., Erez et al., 1996; Raudenbush, 1994). Since it is quite unreasonable in face of the vast literature on the TRA and TPB to assume a common effect size for all studies, the random effects model is used in the present study. All computations followed the procedures as described by Shadish and Haddock (1994) for the integration of effect size estimates.

For moderator analyses we performed weighted regression analyses with effect sizes as dependent and compatibility ratings as independent variables. The weights in these regressions included estimates of random effects variances which had to be estimated in a two-step procedure (method of moments) as detailed in Raudenbush (1994).

15.3 RESULTS

15.3.1 Overall Relationships

In Figure 15.2, overall bivariate relationships for the components of the models are depicted. Except where indicated, results are based on one-factor solutions of principle component analyses. Please note that in the following the effect size measure is the coefficient of determination and not the (multiple) correlation coefficient. Effect sizes might thus look small even though they were quite substantial apart from few exceptions.

In 15 of the studies it was hypothesized that the action under consideration is influenced by factors not under volitional control. The overall effect of only .04% of explained variance shows that perceived behavioral control was

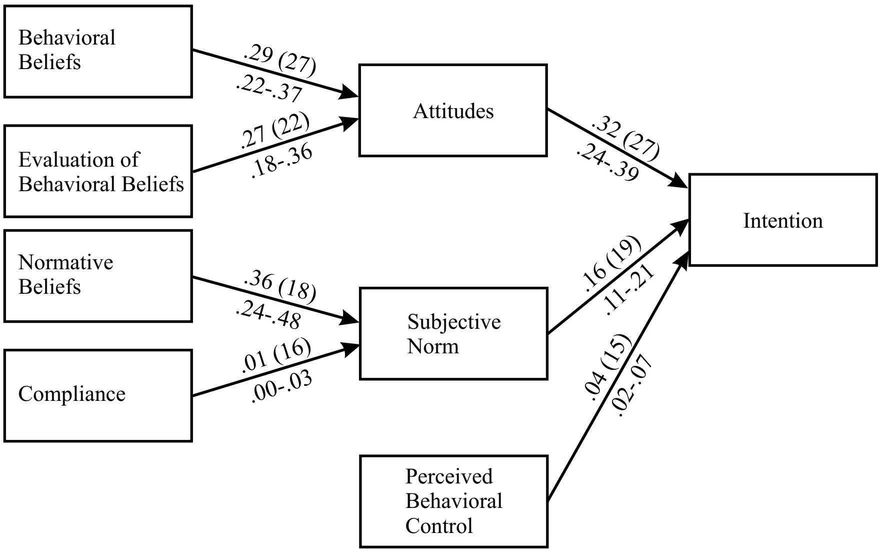


Figure 15.2 Mean effects (R^2) of bivariate relationships (number of studies in brackets and 95% confidence interval below arrows).

not of much importance to predict intention to act, although the 95% confidence interval in Figure 15.2 indicates that this effect is significantly different from zero. To test whether the effect is of importance in the context of attitude and subjective norm, the *F* tests of the individual studies were integrated and revealed no significant incremental variance explained through this component ($p > .05$). The overall mean effect size for the prediction of intention from attitude, subjective norm and perceived behavioral control in 11 studies was .43 with a 95% confidence interval ranging from .35 to .52. The incremental variance explained through subjective norm in the context of attitude was significant ($p < .01$), as might be expected by the predictive power of 16% through subjective norm alone. The overall mean effect size for the prediction of intention from attitude and subjective norm in 19 studies was .38 with a 95% confidence interval ranging from .30 to .45. In sum, as far as the relationship between level II and III components is concerned, strong overall effects on the basis of one-factor solutions were found which are analogous to the bivariate relationships on the basis of unweighted aggregates of items usually reported in applications of the TRA and TPB.

The level III and IV components also showed strong bivariate linear relationships with the exception of subjective norm and compliance. The finding of an absence of a strong effect between these latter components is not unique to the present study but is also reported elsewhere (e.g., Ajzen, 1991, p. 196). Before we provide more details on these relationships in the next subsection, it is interesting to note that the expectancy-value components alone also showed strong bivariate relationships with level III components. The aggregate R^2 for the relationship between the principle component scores of behavioral beliefs and their evaluation expectancy-value products and attitude based on 22 studies was .31 with a 95% confidence interval ranging from .23 to .40. The mean R^2

between subjective norm and the level IV component expectancy-value products was .34 with according confidence interval limits of .27 and .41 on the basis of 16 studies. We will now turn to the results from multiple regressions to assess the incremental value in prediction these single components provide.

15.3.2 Belief Based Measures, Expectancy-Value Components and Multidimensionality

Supplementary to the bivariate results reported, the results from multiple regression of attitude and subjective norm on their antecedent components on level IV of the models are reported in Table 15.4. The values of the homogeneity test based the *Q*-statistic are omitted from the table. They are significant for all the relationships reported in the present study.

Table 15.4 Mean Effects (R^2) and 95% Confidence Intervals for Overall Relationships

Relationship	Mean effect (N)	95% confidence interval
Attitude – BB + EBB	.40 (22)	.31 - .49
Attitude – BB + EBB + EV	.43 (22)	.35 - .51
Attitude – BB + EBB (multi)	.51 (22)	.45 - .58
Attitude – BB + EBB + EV (multi)	.53 (22)	.46 - .59
Subjective norm – NB + CO	.38 (16)	.32 - .44
Subjective norm – NB + CO + EV	.41 (16)	.34 - .47
Subjective norm – NB + CO (multi)	.47 (16)	.40 - .51
Subjective norm – NB + CO (multi)	.49 (16)	.43 - .54

Note. The number of studies with valid data for the relationships is given in brackets. BB = Behavioral beliefs; EBB = Evaluation of behavioral beliefs; NB = Normative beliefs; CO = Compliance; EV = Expectancy-value product; (multi) = multidimensional representation.

To test the impact of combined behavioral beliefs and evaluation of beliefs on attitude, results on the influence of the components and their expectancy-value product are reported in Table 15.4. One way to combine these components is a simple additive combination, which serves as a baseline to test the additional expectancy-value component. As can be seen, the explanatory variance added through the latter is only three percent in the case of a unidimensional and two percent in the case of a multidimensional representation of the components and can be regarded as negligible. Again, results from hierarchical *F* tests for both representations were integrated and showed no significant effect of the expectancy-value component ($p > .05$) in either case. Analogous results emerged with normative beliefs and compliance as predictors of subjective norm. Here, the influence of compliance alone is not significantly different from zero and the expectancy-value component does not add much variance in the context of subjective norm either ($p > .05$). In contrast to the failure

Table 15.5 Results of Random Effects Moderator Analyses for Compatibility as Moderator

Relationship	B	β	SE	$t(df)$	p
Intention – Attitude	.08	.30	.05	1.61 (25)	.06
Intention – SN	.02	.11	.05	.46 (17)	.32
Intention – PBC	.00	.03	.03	.12 (13)	.45
Attitude – BB	.04	.11	.08	.56 (25)	.29
Attitude – EBB	.02	.09	.04	.41 (20)	.34
Attitude – BB + EBB	.07	.18	.09	.41 (20)	.21
SN – CO	.00	.01	.02	.03 (16)	.49

Note. SN = Subjective norm; PBC = Perceived behavioral control; BB = Behavioral beliefs; EBB = Evaluation of behavioral beliefs; B = unstandardized regression coefficient; β = standardized regression coefficient.

of expectancy-value terms to add much variance in prediction of subsequent components, the impact of a multidimensional representation of beliefs and their evaluations is pervasive. For both the prediction of attitude as well as subjective norm, the increase in mean effect sizes is approximately 10%. As multidimensional representations contain overlapping information with unidimensional ones, no test of significance is available.

15.3.3 The Moderating Effect of Compatibility on the Relationships of Components

Table 15.5 reports the results of moderator analyses under the random effects model with mean compatibility ratings from the three raters as independent variables. The computations were performed according to the procedures detailed in Raudenbush (1994).

Descriptively, all regression coefficients are positive, indicating a relationship between compatibility and effect sizes estimates that would be expected from the principle of compatibility with highly compatible components showing higher effect size estimates, and vice versa. As the significance tests reported in the last two columns of Table 15.5 reveal, none of the relationships is significant according to conventional criteria. For significance tests under the random effects approach it is important to bear in mind that they are more conservative than alternative tests under the fixed effects approach which are mostly applied (Hedges & Vevea, 1998). Furthermore, as the relatively large non-significant coefficients for the relationships between intention and attitude show, the number of studies in the present meta-analysis might not be sufficient to achieve high levels of statistical power. In addition to the estimation and tests of regression parameters, the residual variances after taking the predictors into account were tested for significance. These analyses revealed that for all relationships reported in Table 15.5 significant variances remained to be explained.

Table 15.6 Results for Attitude Toward Object vs. Attitude Toward Behavior

	Attitude toward behavior		Attitude toward object	
Relationship	Mean effect (N)	95% confidence interval	Mean effect (N)	95% confidence interval
Intention – Attitude	.38 (19)	.32 - .44	.17 (8)	.01 - .33
Attitude – BB	.32 (19)	.23 - .40	.24 (8)	.07 - .42
Attitude – EBB	.27 (17)	.17 - .37	.27 (5)	.03 - .50
Attitude – BB + EBB	.41 (17)	.32 - .49	.38 (5)	.11 - .67

Note. The number of studies with valid data for the relationships is given in brackets. BB = Behavioral beliefs; EBB = Evaluation of behavioral beliefs.

An alternative classification for high vs. low compatibility groups in the context of attitude assessment was undertaken following the suggestions of Eckes and Six (1994). They argued that following the principle of compatibility attitude toward an object is always less compatible to other components than attitude toward behavior, because the action element is missing and the other dimensions of compatibility are usually left unspecified. The mean effect sizes for a comparison of these groups are reported in Table 15.6.

The results replicate the findings of Eckes and Six (1994) that attitude toward an object showed lower relationships with other components in the models than attitude toward behavior. Despite this clear trend of decline of explained variance for low compatibility groups, the confidence intervals for all effects were again overlapping, so overall the differences between these two groups were not significant. Additionally, it should be emphasized that the tests for homogeneity in all groups were still significant, thereby calling for more or alternative moderators to explain observed variances in effect sizes within these groups. In sum, for both approaches the compatibility between model components showed consistent but non-significant and partly small effects as predictors of effect size variance.

15.4 DISCUSSION AND CONCLUSIONS

The present study investigated the relationships of the components of the TRA and TPB in a series of hitherto unpublished studies. For these studies would not have been published without the present study, this can be regarded as a "grasp into the file-drawer" (Rosenthal, 1979). In contrast to the expectations of critics of meta-analysis, strong effects were found in the file-drawer. The results of the present study fit well within the context of other meta-analyses (see Section 15.1.2), thereby re-emphasizing the importance of attitude as a psychological construct for the explanation and prediction of behavior and the utility of the TRA and TPB in general. For there are no remarkable differences between the effects of published meta-analyses and the results reported here, this is interpreted as support for the hypothesis that the unpublished studies under review do not markedly differ from published studies. It might nevertheless be suspected that these studies, though not different in effects, are characterized by other features that serve as alternative explanations of the effects reported. Here, it might be added and reiterated that first, the persons who conducted the studies were not aware of the fact that a meta-analysis will be performed on their data at the time of conducting their study. Second, they chose their field of application at their own discretion and were only influenced by the second author of the present article as to make them follow the recommendations by Ajzen and Fishbein (1980). This was reviewed during realization of the studies and in the secondary analyses. Third, as might be suspected, this influence did not result in extremely homogeneous study effects. To the contrary, effect size variances were all significant, even under the random effects model. Fourth, the raters of compatibility were not aware of any result of the studies, so a potential influence by this knowledge influence was precluded. In sum, we argue that it might be implausible to attribute our findings to special characteristics of our study sample.

One of our findings was that PBC has not emerged as an important determinant of intentions to act. This might be due to several possible causes. As has been indicated, the reliability of this component was lower than the reliability of all other components and this might have contributed to a reduced relationship with intention. Next, PBC may be an important predictor in our studies for behavior but not for intentions, although this is somewhat implausible against the background of the results referred to in Section 15.1.2. In face of these mixed results it is not warranted to renounce perceived behavioral control as a predictor of intentions or behavior but it obviously did not always have an influence on intentions when expected by the primary researchers. Therefore, we agree with Petty et al. (1997) in that research that goes beyond speculations of the influence of contextual factors is needed to clarify circumstances under which PBC is an essential predictor.

Another finding of the present study was that expectancy-value components did not add significantly to the prediction of subsequent components, an aspect not considered in previous meta-analyses. Incremental variance explained not only was insignificant, the magnitude of the effect was also quite small. As a result, one is left with a good prediction model consisting of an additive combination of level IV components which might not make much sense in psychological terms (Eagly & Chaiken, 1993). The difficult situation here is that psychologically meaningful scaling of belief items results in psychometrically meaningless or arbitrary correlations with other components, while proper methods from the viewpoint of measurement theory may lead to psychologically meaningless results (Bagozzi, 1984). Although this difficult subject has been addressed quite often (e.g., Orth, 1986; Sparks et al., 1991), it is not recognized by all primary researchers (Evans, 1991).

In addition, the issue of multidimensionality of belief structures has been of special concern in the present study. It was also Bagozzi who pointed out that "If people at times form multidimensional attitudes or if one desires to learn which beliefs and evaluations are most important, then the Fishbein model may not be useful and may even mislead the researcher" (Bagozzi, 1984, p. 301). The results from the present study underscore the importance of this issue. In all twenty-two studies, which provided data for behavioral beliefs and their evaluations, it was impossible to determine at least two different meaningful dimensions and these contributed substantially to the prediction of the attitude component. This calls into question the assumption of unidimensional belief structures, leading to both a better prediction and explanation model of attitudes.

But what are the costs and benefits of representing the level IV components as multidimensional in general? It is admitted that parsimony of the TRA and TPB may be regarded as sacrificed for a questionable gain of enhanced prediction. Even the danger of excessive "data fitting" may be seen in an approach that advocates the exploration of multidimensional structures. To be clear, it is not advisable to subdivide level IV variables in as much components as possible. We instead propose to explicitly test measurement models for all components of the model where possible. Only in cases where a multidimensional structure clearly emerges and is theoretically sensible there is the potential to enhance prediction and, at least as equally important, understanding of the formation of components on different levels of the model. These benefits are achieved through the specification of distinguishable dimensions in the domains of behavioral consequences, normative influences, and control. Moreover, these dimensions are tested for their differential impact on other components of the model by estimation of the dimensional weights so that the formation of attitude in a particular application, for example, can be more clearly traced back to specific antecedents. How these weights are to be interpreted is not definitely clear yet. One possible interpretation is that they represent importance weights of the dimensions for the formation of attitudes (for a review on this issue, see van der Pligt, de Vries, Manstead, & van Harreveld, 2000). That is, these weights can be interpreted as an "empirical filter" for characteristics represented in the items of level IV components that are not predictive (or important) of attitudes. Another possible interpretation is that the weights function to pronounce more accessible dimensions in contrast to less accessible dimensions. In either way, the empirical results reported by van der Pligt et al. (2000) that items selected for importance correlate more highly with attitude and behavior/intention than nonselected items is in accordance with our results and lends support to the notion of these weights as importance factors. Indeed, in most but not all cases, weights for the multiple dimensions on level IV were not all significant but variance explained in attitudes increased in all cases, even as measured by adjusted R^2. Unfortunately, we could not integrate these results in our meta-analyses for technical reasons, so we did only report them here descriptively.

A final benefit of a multidimensional representation is that it offers the possibility to assess whether an inconsistent belief basis may exist or even prevail in a certain context. Such an inconsistent belief basis can result in attitudinal ambivalence at least for some persons, a phenomenon of attitude structure that is known in attitude research for quite a long time (Scott, 1966, 1969) and has attracted remarkable research activities in recent years (e.g., Cacioppo, Gardner, & Berntson, 1997; Jonas, Diehl, & Brömer, 1997). Since attitudinal ambivalence has also been shown to moderate the attitude–behavior relationship (Jonas et al., 1997), the exploration of multidimensional belief structures seems to be a useful tool to assess whether attitudinal ambivalence is of relevance in a given study. In our view, inconsistencies of beliefs are not limited to behavioral beliefs but may also occur with normative beliefs.

The second major issue of the present study was testing the principle of compatibility as a moderator in applications of the TRA and TPB where we extended the application of this principle to all model relationships. Most of the previous meta-analyses in Table 15.1 attempted to account for observed variability in effect sizes but there has not yet emerged a small set of moderators potent enough to give an explanation of this variability. Nearly all of the attempts to account for variability – like the present study – focused on seemingly methodological explanations of which the principle of compatibility seemed to be the most interesting one, because it was supposed to give an answer to the challenge of attitude as a psychological construct put forward by Wicker (1969). The present study showed that indeed part of the variability of effect sizes in the TRA and TPB could be explained by differences between studies concerning compatibility of components, but overall, the explanatory effect of compatibility was somewhat low and disappointing. This result may indicate that the principle does not necessarily work with the force ascribed to it or that it does not do so for all relationships of the TRA and TPB. Whereas initially the principle of compatibility was confined to a methodological characteristic, it has recently been tied to more psychologically meaningful interpretations (Ajzen & Sexton, 1999). The authors argue that if beliefs accessed in the attitudinal and behavioral context are the same, high correlations can be expected. This match in beliefs might be facilitated through a match of components on the TACT-dimensions, although they note that biases in belief elicitation in the different contexts can also lead to low correlations despite highly compatible components. Tracing the roots of the principle of compatibility down to belief congruence and linking it to theoretical approaches like the attitude representation theory (Lord & Lepper, 1999) seems to be a promising approach for further research because it illuminates how the principle actually works in psychological terms and when it may fail to work.

REFERENCES

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action control. From cognition to behavior* (pp. 1–39). Berlin: Springer.

Ajzen, I. (1988). *Attitude, personality, and behavior.* Milton Keynes: Open University Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211.

Ajzen, I. (1996). The directive influence of attitudes on behavior. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 385–403). New York: Guilford Press.

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*, 888–918.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs: Prentice-Hall.

Ajzen, I., & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavior control. *Journal of Experimental Social Psychology*, *22*, 453–474.

Ajzen, I., & Sexton, J. (1999). Depth of processing, belief congruence, and attitude–behavior correspondence. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 117–138). New York: Guilford.

Allport, G. W. (1935). Attitudes. In C. Murchinson (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Armitage, C. J., & Conner, M. (1999). The theory of planned behavior: Assessment of predictive validity and 'perceived control'. *British Journal of Social Psychology*, *38*, 35–54.

Bagozzi, R. P. (1981a). Attitudes, intention, and behavior: A test of some key hypotheses. *Journal of Personality and Social Psychology*, *41*, 607–627.

Bagozzi, R. P. (1981b). An examination of the validity of two models of attitude. *Multivariate Behavioral Research*, *16*, 323–359.

Bagozzi, R. P. (1984). Expectancy–value attitude models: An analysis of critical measurement issues. *International Journal of Research in Marketing*, *1*, 295–310.

Bagozzi, R. P., & Kimmel, S. K. (1995). A comparison of leading theories for the prediction of goal-directed behaviors. *British Journal of Social Psychology*, *34*, 437–461.

Blue, C. L. (1995). The predictive capacity of the theory of reasoned action and the theory of planned behavior in exercise research: An integrated literature review. *Research in Nursing and Health*, *18*, 105–121.

Brown, S. P., & Stayman, D. M. (1992). Antecedents and consequences of attitude toward the ad: A meta-analysis. *Journal Of Consumer Research*, *19*, 34–51.

Burnkrant, R. E., & Page, T. J. (1988). The structure and antecedents of the normative and attitudinal components of Fishbein's theory of reasoned action. *Journal of Experimental Social Psychology*, *24*, 66–87.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*, 3–25.

Cliff, N. (1988). The eigenvalues–greater–than–one rule and the reliability of components. *Psychological Bulletin*, *103*, 276–279.

Cliff, N., & Caruso, J. C. (1998). Reliable component analysis through maximizing composite reliability. *Psychological Methods*, *3*, 291–308.

Conner, M., & Armitage, C. J. (1998). Extending the theory of planned behavior: A review and avenues for further research. *Journal of Applied Social Psychology*, *28*, 1429–1464.

Conner, M., & Sparks, P. (1996). The theory of planned behaviour and health behaviours. In M. Conner & P. Norman (Eds.), *Predicting health behaviour: Research and practice with social cognition models* (pp. 121–162). Buckingham: Open University Press.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Orlando, FL: Harcourt Brace.

Eagly, A. H., & Chaiken, S. (1998). Attitude structure and function. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (Vol. 2, pp. 269–322). Boston, MA: McGraw-Hill.

Eckes, T., & Six, B. (1994). Fakten und Fiktionen in der Einstellungs-Verhaltens-Forschung: Eine Meta-Analyse [Facts and fiction in attitude–behavior research: A meta-analysis]. *Zeitschrift für Sozialpsychologie*, *25*, 253–271.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, *49*, 275–306.

Evans, M. G. (1991). The problem of analyzing multiplicative composites. *American Psychologist*, *46*, 6–15.

Farley, J. U., Lehman, D. R., & Ryan, M. J. (1981). Generalizing from "imperfect" replication. *Journal of Business*, *54*, 597–610.

Fishbein, M. (1963). An investigation of the relationships between beliefs about an object and the attitude toward that object. *Human Relations*, *16*, 233–240.

Fishbein, M. (1967). A consideration of beliefs, and their role in attitude measurement. In M. Fishbein (Ed.), *Readings in attitude measurement and theory* (pp. 257–266). New York: Wiley.

Fishbein, M. (1980). A theory of reasoned action: Some applications and implications. In M. M. Page (Ed.), *Nebraska symposium on motivation 1979* (pp. 65–116). Lincoln: University of Nebraska Press.

Fishbein, M., & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review*, *81*, 59–74.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Godin, G., & Kok, G. (1996). The theory of planned behavior: A review of its applications to health-related behaviors. *American Journal of Health Promotion*, *11*, 87–98.

Gorsuch, R. L., & Ortberg, J. (1983). Moral obligation and attitudes: Their relation to behavioral intentions. *Journal of Personality and Social Psychology*, *44*, 1025–1028.

Grube, J., Morgan, M., & McGree, S. T. (1986). Attitudes and normative beliefs as predictors of smoking intentions and behavior: A test of three models. *British Journal of Social Psychology*, *25*, 81–93.

Hausenblas, H. A., Carron, A. V., & Mack, D. E. (1997). Application of theories of reasoned action and planned behavior to exercise behavior: A meta-analysis. *Journal of Sport & Exercise Psychology*, *19*, 36–51.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388–395.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* London: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Jaccard, J. J., King, W., & Pomazal, R. (1977). Attitudes and behavior: An analysis of specificity of attitudinal predictors. *Human Relations*, *30*, 817–824.

Jonas, K., Diehl, M., & Brömer, P. (1997). Effects of attitudinal ambivalence on information processing and attitude–intention consistency. *Journal of Experimental Social Psychology*, *33*, 190–210.

Jonas, K., & Doll, J. (1996). Eine kritische Bewertung der Theorie überlegten Handels und der Theorie geplanten Verhaltens [A critical evaluation of the theory of reasoned action and the theory of planned behavior]. *Zeitschrift für Sozialpsychologie*, *27*, 18–31.

Kim, M. S., & Hunter, J. E. (1993). Attitude–behavior relations: A meta-analysis of attitudinal relevance and topic. *Journal of Communication*, *43*, 101–142.

Kirkpatrick, L. A. (1997). Effects of multiple determinacy and measurement error on trait–behavior and behavior–behavior relations: An integrated conceptual model. *Personality and Social Psychology Bulletin*, *23*, 199–209.

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin*, *21*, 58–75.

LaPiere, R. T. (1934). Attitude vs. actions. *Social Forces*, *13*, 230–237.

Lord, C. G., & Lepper, M. R. (1999). Attitude representation theory. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 31, pp. 265–343). San Diego, CA: Academic Press.

Manstead, A. S. R. (2000). The role of moral norm in the attitude–behavior relation. In D. J. Terry & M. A. Hogg (Eds.), *Attitudes, behavior, and social context: The role of norms and group membership* (pp. 11–30). Mahwah, NJ: Lawrence Erlbaum.

Millar, M. G., & Tesser, A. (1986). Effects of affective and cognitive focus on the attitude–behavior relation. *Journal of Personality and Social Psychology*, *51*, 270–276.

Millar, M. G., & Tesser, A. (1992). The role of beliefs and feelings in guiding behavior: The mismatch model. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgements* (pp. 277–339). Hillsdale, NJ: Lawrence Erlbaum.

Nesselroade, J. R., & McArdle, J. J. (1997). On the mismatching of levels of abstraction in mathematical-statistical model fitting. In H. W. Reese & M. D. Franzen (Eds.), *Life-span developmental psychology: Biological and neuropsychological mechanisms* (pp. 23–39). Hillsdale, NJ: Lawrence Erlbaum.

Notani, A. S. (1998). Moderators of perceived behavioral control's predictiveness in the theory of planned behavior: A meta-analysis. *Journal of Consumer Psychology*, *7*, 247–271.

Orth, B. (1986). Messtheoretisch bedeutsame oder psychologisch sinnvolle Einstellungsmodelle [Attitude models meaningful for measurement theory or of psychological significance]. *Zeitschrift für Sozialpsychologie*, *17*, 87–90.

Parker, D., Manstead, A. S. R., & Stradling, S. (1995). Extending the theory of planned behavior: The role of personal norm. *British Journal of Social Psychology*, *34*, 127–137.

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology*, *48*, 609–647.

Randall, D. M., & Wolff, J. A. (1994). The time interval in the intention–behavior relationship: Meta-analysis. *British Journal of Social Psychology*, *33*, 405–418.

Raudenbush, S. W. (1994). Random effects models. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Ryan, M., & Bonfield, E. H. (1975). The Fishbein extended model and consumer behavior. *Journal of Consumer Research*, *2*, 118–136.

Schlegel, R. P., & DiTecco, D. (1982). Attitudinal structures and the attitude–behavior relation. In M. P. Zanna, E. T. Higgins, & C. P. Herman (Eds.), *Consistency in social behavior: The Ontario symposium* (Vol. 2, pp. 17–49). Hillsdale, NJ: Lawrence Erlbaum.

Scott, W. A. (1966). Measures of cognitive structure. *Multivariate Behavioral Research*, *1*, 391–395.

Scott, W. A. (1969). Structure of natural cognitions. *Journal of Personality and Social Psychology*, *12*, 261–278.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Sheeran, P., & Taylor, S. (1999). Predicting intentions to use condoms: A meta-analysis and comparison of the theories of reasoned action and planned behavior. *Journal of Applied Social Psychology*, *29*, 1624–1675.

Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, *15*, 325–343.

Sherman, S., & Fazio, R. H. (1983). Parallels between attitudes and traits as predictors of behavior. *Journal of Personality*, *51*, 308–345.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Six, B., & Eckes, T. (1996). Metaanalysen in der Einstellungs-Verhaltens-Forschung [Meta-analyses in attitude–behavior research]. *Zeitschrift für Sozialpsychologie*, *27*, 7–17.

Sparks, P. (2000). Subjective expected utility-based attitude–behavior models: The utility of self-identity. In D. J. Terry & M. A. Hogg (Eds.), *Attitudes, behavior, and social context: The role of norms and group membership* (pp. 31–46). Mahwah, NJ: Lawrence Erlbaum.

Sparks, P., Guthrie, C. A., & Shepherd, R. (1997). The dimensional structure of the perceived behavioral control construct. *Journal of Applied Social Psychology*, *27*, 418–438.

Sparks, P., Hedderly, D., & Shepherd, R. (1991). Expectancy-value models of attitudes: A note on the relationship between theory and methodology. *European Journal of Social Psychology*, *21*, 261–271.

Sparks, P., Shepherd, R., & Frewer, L. J. (1995). Assessing and structuring attitudes toward the use of gene technology in food production: The role of perceived food production: The role of perceived ethical obligation. *Basic and Applied Social Psychology*, *16*, 267–285.

Stroebe, W., Eagly, A. H., & Ajzen, I. (1996). Individuelle Unterschiede im Verhalten: Das sozialpsychologische Forschungsprogramm [Interindividual differences in behavior: The research program of social psychology]. In K. Pawlik (Ed.), *Grundlagen und Methoden der differentiellen Psychologie* (pp. 242–267). Göttingen: Hogrefe.

Sutton, S. (1998). Predicting and explaining intentions and behavior: How well are we doing? *Journal of Applied Social Psychology*, *28*, 1317–1338.

Terry, D., & O'Leary, J. E. (1995). The theory of planned behavior: The effects of perceived behavioural control and self-efficacy. *British Journal of Social Psychology*, *34*, 199–220.

Triandis, H. C. (1980). Values, attitudes, and interpersonal behavior. In H. E. Howe, Jr. & M. M. Page (Eds.), *Nebraska symposium on motivation, 1979* (Vol. 27, pp. 195–259). Lincoln: University of Nebraska.

van den Putte, B. (1991). *20 years of the theory of reasoned action of Fishbein and Ajzen: A meta-analysis.* Unpublished manuscript, University of Amsterdam, Netherlands.

van der Pligt, J., de Vries, N. K., Manstead, A. S. R., & van Harreveld, F. (2000). The importance of being selective: Weighing the role of attribute importance in attitudinal judgment. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 135–200). New York: Academic Press.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*, 41–78.

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 505–560). New York: Plenum Press.

Acknowledgments

We would like to thank the following individuals for their invaluable help in assessing the compatibility ratings of the studies: Tina Block, Christian Ehrlich, Jens Tangenberg, Michaela Turß, Raphael Vogel, and Claudia Winter. We are also grateful to the persons who conducted the original 27 studies and made the data sets available to us.

16

META – A Software Package for Meta-Analysis in Medicine, Social Sciences, and the Pharmaceutical Industry

Peter Schlattmann
Uwe Malzahn
Dankmar Böhning

Working Group: Biometry and Epidemiology
Institute for International Health, Joint Center for Humanities and Health Sciences
Humboldt-University in Berlin / Free University Berlin

Summary

The software META provides statistical methods for the performance of meta-analyses in medicine, psychology, and quality assurance in the pharmaceutical industry. META makes a variety of effect measures available, like the relative risk, the standardized difference, and quality indices. For these effect measures, classical pooled estimators as well as “modern” random effect models can be calculated, for example, the approach of DerSimonian and Laird (1986) or the mixture distribution approach (Böhning, 2000a; Böhning et al., 1998). The latter approach allows the semi-parametric estimation of the heterogeneity structure and classification of individual studies or batches. In addition to statistical methods there are graphical facilities, such as funnel plots for the identification of a publication bias or plots of confidence intervals for an illustration of individual studies and the pooled effect measure. META is a public domain program. It comes with a graphical interface and is available for Windows 9x/NT and Unix (Linux).

16.1 INTRODUCTION

In the past few years, meta-analysis has become increasingly popular in many areas of science such as medicine, psychology, and other social sciences. In these areas of application meta-analyses have been performed in order to obtain a pooled estimate of various single studies. Obtaining a single summary measure implicitly assumes *homogeneity* of these studies, that is, the results of individual studies differ only by chance. In this case a combined estimate of the individual studies provides a powerful and important result. However, this pooled estimate may be seriously misleading if study conditions are *heterogenous*.

Thus, an approach which considers meta-analysis as a study over studies has increasingly been advocated. This approach seeks to investigate heterogeneity between studies. An important feature of this type of meta-analysis lies in the fact that it tries to identify factors which cause heterogeneity.

This approach may easily be extended to the area of quality control, where batches of the produced goods replace the role of studies in medicine or the social sciences. Clearly, in this setting an investigation of heterogeneity is equally attractive, since identification and modeling of heterogeneity helps to improve the production process. An introduction how to use the methodology of meta-analysis in quality control is given by Böhning and Dammann in Chapter 10 of this volume.

16.2 THE PROGRAM META

The software META has been developed to provide a tool which allows to perform meta-analyses within the areas of application described above. The focus of META is on the analysis of heterogeneity, which may be considered here the unifying concept for several fields of application.

For different areas of application, different measures of effects are important and necessary. Thus, META enables the meta-analyst to choose out of a variety of measures of effects, such as the relative risk in medicine, the standardized difference in psychology and proportions in quality control, just to mention a few.

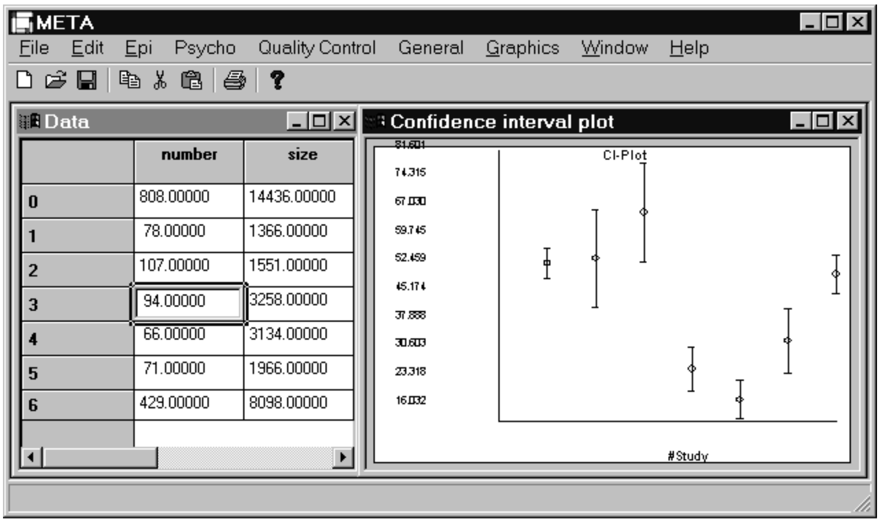
META provides various statistical methods to perform meta-analyses such as simple pooled estimates, random effects models, and graphical procedures such has confidence interval plots, funnel plots, and so forth. We will illustrate the possible use of META using a data set from psychiatric epidemiology.

16.3 A WORKED EXAMPLE

The following meta-analysis investigates the prevalence of agoraphobia based on seven studies (Eaton, 1995) in several countries all over the world. Agoraphobia may be defined as space anxiety, as a fear of being in public places.

This psychiatric disorder may even lead to total avoidance of public places and thus may cause severe disability.

An initial step in any meta-analysis might be to plot the effect measure together with a 95% confidence interval. This may be done using META and its graphics facilities. Figure 16.1 shows a screen dump of META and its data window. The data window shows the prevalent cases of agoraphobia together with the population at risk of the respective study.



	number	size
0	808.00000	14436.00000
1	78.00000	1366.00000
2	107.00000	1551.00000
3	94.00000	3258.00000
4	66.00000	3134.00000
5	71.00000	1966.00000
6	429.00000	8098.00000

Figure 16.1 Data window and confidence interval plot.

The simplest model possible assumes parametric density $f(x, \theta, \sigma^2)$ for some random quantity X where θ is a parameter of interest and σ^2 is a nuisance parameter which might or might not be present in the model. In the example at hand, $f(x, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. In this case all studies are assumed to measure the same overall effect θ, and they only differ in variability. Thus, the summary measure needs to assign weights according to the inverse of the variance of the individual study in order to obtain the summary measure.

Looking at the confidence interval plot, there seems to be a large degree of variability to be present. However, frequently one is interested in obtaining a summary measure for all studies. Using META we obtain the following results:

```
POOLED ESTIMATOR FOR PROPORTIONS

RESULTS
Pooled estimate:     0.048892
Common variance:   0.00000145

95 percent confidence interval (0.04654, 0.05125)

Chi-Square test for homogeneity of proportions:
     115.23539 df = 6  p-value:      0.00000
```

Clearly, looking at the value of the χ^2 test of homogeneity, we reject the null-hypothesis and conclude that there is substantial heterogeneity in terms of the prevalence of agoraphobia in the countries studied. As a result, the computation of an overall rate is not very meaningful, since we would ignore the underlying heterogeneity.

In order to deal with heterogeneity, a two-level model is implemented in META. As before, $f(x, \theta, \sigma^2)$ denotes a parametric density for some random quantity X. But now it is assumed that θ is not constant but is varying itself according to some further distribution P for which the moments $E_P(\theta) = \mu$ and $Var_P(\theta) = \tau^2$ are assumed to exist. Consequently, we are lead to a *marginal* or *unconditional* distribution $f(x, P) = \int f(x, \theta) P(d\theta)$.

Frequently, τ^2 is called the *heterogeneity variance*. META offers modeling according to two different distributions in order to deal with heterogeneity: one is the moment approach which is based on equating the expected value of the χ^2-statistic to the observed one and then solving for τ^2. Actually, this is the approach by DerSimonian and Laird (1986). The other approach does not specify P any further and leads to the marginal density, a mixture model. Here, $f(x, P) = \sum_{j=1}^{k} p_j f(x_i, \theta_j, \sigma_i^2)$. According to this model, we assume the existence of k subpopulations with parameters θ_j receiving weight p_j for the j^{th} subpopulation. A detailed description of the use of this approach in meta analysis may be found in Böhning et al. (1998), or in Böhning (2000a).

We proceed in our analysis with the estimation of the DerSimonian-Laird estimator:

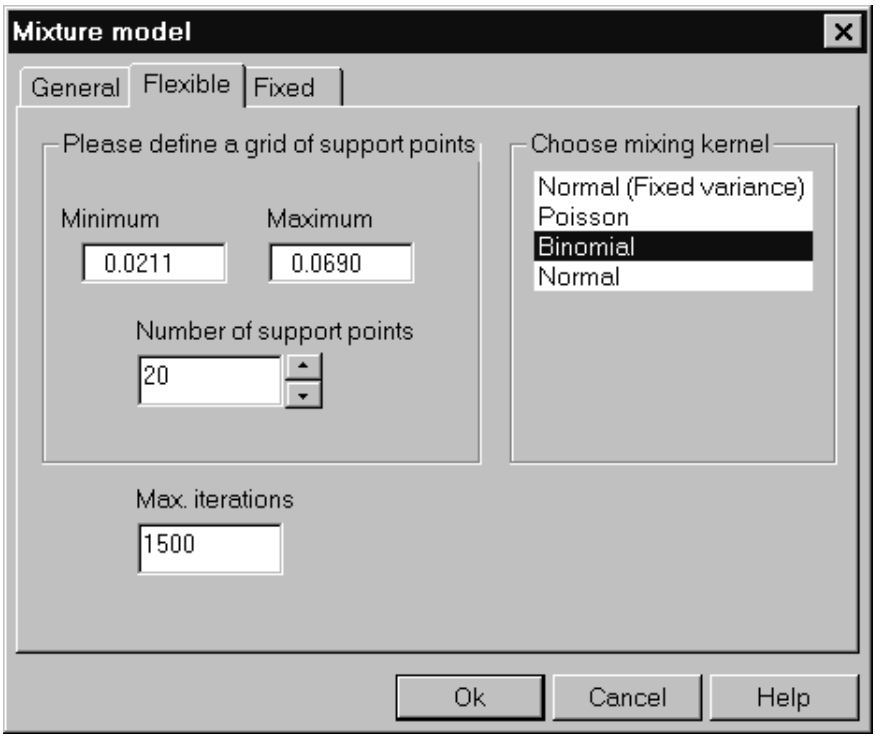
```
RESULTS
Pooled DerSimonian-Laird estimate:       0.0455

Heterogeneity variance:       0.0003

Variance of pooled estimator: 0.0000465

0.04545 95 percent CI: (0.0321,   0.0588)
```

Please note that we find a substantial value for the heterogeneity variance τ^2 in this data set. As expected, incorporating heterogeneity leads to a larger variance for the DerSimonian-Laird estimator. As a result, we obtain a much wider confidence interval compared to the pooled estimator where we assume a constant value for θ.

Frequently, there is a debate whether one should use a summary measure in the presence of heterogeneity. One might argue that this may be done, but one has to be careful how to interpret the results. Under the presence of heterogeneity a summary measure will reflect the overall mean in the population well, knowing that this effect might be different in subparts of the population.

If the presence of heterogeneity has been identified, one might wish to model the *structure* of this heterogeneity and, for example, find the levels of effect in subparts of the population. This can be accomplished using the finite mixture model approach outlined above. A convenient computational strategy uses a fixed grid of potential support points (subpopulation means θ_j) which may or may not receive weights p_j.

Figure 16.2 shows the dialog box which allows the user to define a grid of potential support points.



Please define a grid of support points

Minimum: 0.0211

Maximum: 0.0690

Number of support points: 20

Choose mixing kernel: Normal (Fixed variance) / Poisson / Binomial / Normal — Binomial selected

Max. iterations: 1500

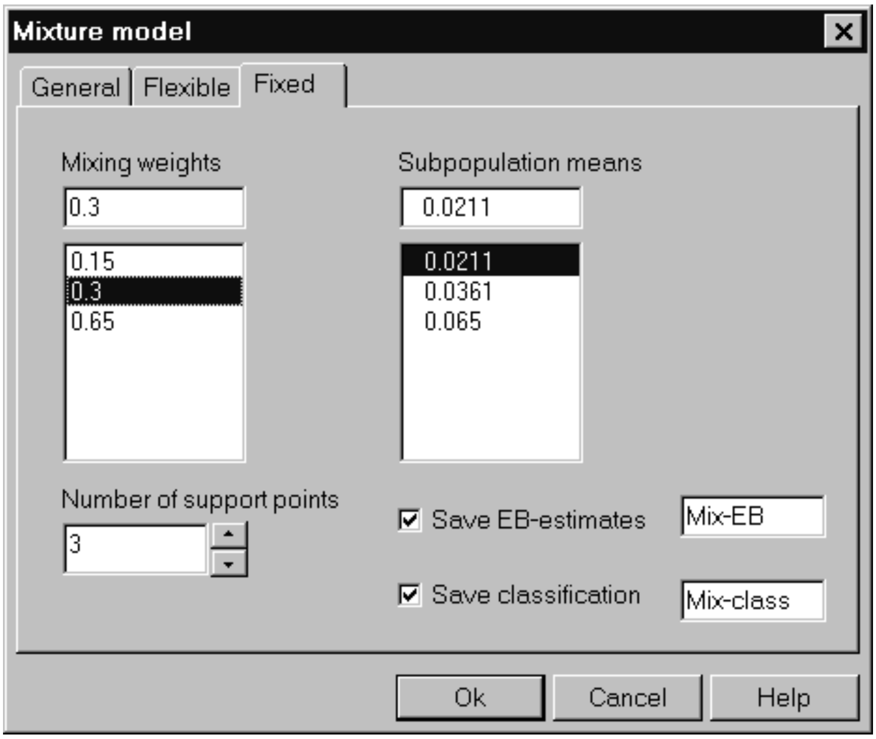
Figure 16.2 Dialog box for the definition of a grid of potential support points in the mixture model.

Depending on the current measure of effect an appropriate mixing kernel may be chosen by the user. In this case – since we are dealing with rates – the binomial distribution is the natural choice.

```
Initial number of components: 5
Parameter:   0.0211, Weight:   0.1441
Parameter:   0.0317, Weight:   0.2840
Parameter:   0.0530, Weight:   0.3073
Parameter:   0.0584, Weight:   0.1533
Parameter:   0.0690, Weight:   0.1113

Log-likelihood at iterate: -34.8009
```

Based on this grid META identifies five potential subpopulations. Now these grid points with positive support may be used to find a refined solution using the EM-algorithm (Dempster et al., 1977). Here, we keep the number of components fixed and update mixing weights and subpopulation means. Frequently, some population means coincide and thus the number of components decreases. For our data at hand, after applying the EM-algorithm, we find four remaining components (results not shown here).

Now a backward elimination approach may be used in order to reduce the number of mixing components. This would imply that we test k = 4 vs. k =3 using a Likelihood Ratio test approach (see Figure 16.3).



Mixing weights: 0.3
0.15
0.3
0.65

Subpopulation means: 0.0211
0.0211
0.0361
0.065

Number of support points: 3

☑ Save EB-estimates: Mix-EB

☑ Save classification: Mix-class

Figure 16.3 Dialog box for fixed effect mixture model.

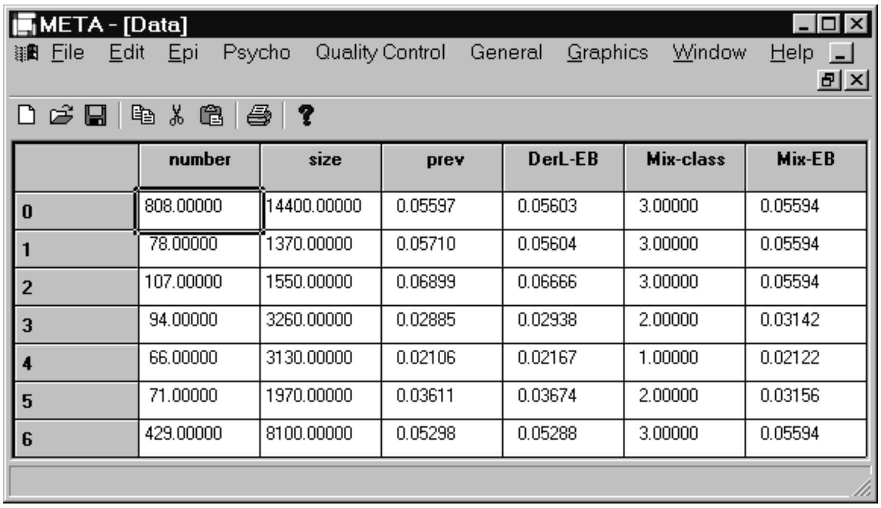
```
NPMLE for Fixed support size

Number of components after combining equal parameter estimates: 3

Parameter:   0.0212, Weight:   0.1440
Parameter:   0.0316, Weight:   0.2844
Parameter:   0.0559, Weight:   0.5716

Log-likelihood at iterate: -34.3889
```

Clearly, the log-likelihood is only slightly smaller for this three component mixture model and we would conclude that a three component solution is appropriate. Once a mixture model has been chosen, one might be interested in classifying the individual study. Due to their discrete structure, mixture models provide a natural way of classifying the individual study. This is achieved by applying Bayes theorem and using the estimated mixing distribution as a prior distribution. Thus, we are able to compute the posterior probability for each study to belong to a certain component:

$$Pr(Z_{ij} = 1 | x_i, \hat{P}) = \frac{\hat{p}_j f(x_i, \hat{\theta}_j)}{\sum\limits_{l=1}^{k} \hat{p}_l f(x_i, \hat{\theta}_l)}.$$

The *i*th study is then assigned to that subpopulation *j* for which it has the highest posterior probability of belonging. META offers the option to classify the studies and to store the results of this classification in the data spreadsheet (see Figure 16.3).

META also computes the posterior expectation for the measure of effect for the individual study based on the assumed distribution. Likewise, the posterior expectations may also be stored within the data frame as may be seen in Figure 16.4.



	number	size	prev	DerL-EB	Mix-class	Mix-EB
0	808.00000	14400.00000	0.05597	0.05603	3.00000	0.05594
1	78.00000	1370.00000	0.05710	0.05604	3.00000	0.05594
2	107.00000	1550.00000	0.06899	0.06666	3.00000	0.05594
3	94.00000	3260.00000	0.02885	0.02938	2.00000	0.03142
4	66.00000	3130.00000	0.02106	0.02167	1.00000	0.02122
5	71.00000	1970.00000	0.03611	0.03674	2.00000	0.03156
6	429.00000	8100.00000	0.05298	0.05288	3.00000	0.05594

Figure 16.4 Spreadsheet with original data and empirical Bayes estimates.

16.4 AVAILABILITY

META is designed to be platform independent and uses the wxWindows 2.0 class library (Smart, 2000). META may be obtained for Microsoft Windows 9x/NT and for Unix(Linux) operating systems. META is available from the authors on request.

REFERENCES

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping, and others.* Boca Raton, FL: Chapman & Hall/CRC.

Böhning, D., Dietz, E., & Schlattmann, P. (1998). Recent developments in computer-assisted analysis of mixtures. *Biometrics*, *54*, 525–536.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Eaton, W. W. (1995). Progress in the epidemiology of anxiety disorders. *Epidemiologic Reviews*, *17*, 32–38.

Smart, J. (2000). wxwindows – a cross-platform GUI-solution. Available via WWW: http://www.wxwindows.org.

Contributors

DR. DOROTHEE ABRAHAM-RUDOLF, Psychiatrische Klinik der Technischen Universität München, Ismaninger Str. 22, D-81675 München

DR. GERD ANTES, Deutsches Cochrane Zentrum, Universitätsklinikum der Albert-Ludwigs-Universität Freiburg, Institut für Medizinische Biometrie und Medizinische Informatik, Stefan-Meier-Str. 26, D-79104 Freiburg; antes@cochrane.de

DR. DOGAN ARGAC, Fachbereich Statistik, Universität Dortmund, Vogelpothsweg 87, D-44221 Dortmund; argac@statistik.uni-dortmund.de

PROF. DR. MARIA BLETTNER, Epidemiologie und Medizinische Statistik, Fakultät für Gesundheitswissenschaften, Universität Bielefeld, Postfach 10 01 31, D-33501 Bielefeld; blettner@uni-bielfeld.de

PROF. DR. DANKMAR BOEHNING, AG Biometrie und Epidemiologie, Institut für Soziale Medizin, Zentrum für Human- und Gesundheitswissenschaften, Freie Universität Berlin / Humboldt-Universität zu Berlin, Fabeckstr. 60-62, D-14195 Berlin; boehning@zedat.fu-berlin.de

MICHAELA BROCKE, Psychologisches Institut IV, Westfälische Wilhelms-Universität, Fliednerstr. 21, D-48149 Münster; brockem@psy.uni-muenster.de

DR. UWE CZIENSKOWSKI, Max Planck Institute for Human Development, ABC Research Group, Lentzealle 94, D-14195 Berlin; sciencec@zedat.fu-berlin.de

DR. SUSANNE DAHMS, Institut für Biometrie und Informationsverarbeitung, Freie Universität Berlin, Oertzenweg 19b, D-14163 Berlin; sdahms@zedat.fu-berlin.de

UWE-PETER DAMMANN, Asta Medica AG, Kantstr. 2, D-33790 Halle / Westfalen; Uwe-Peter.Dammann@astamedica.de

DR. EKKEHART DIETZ, AG Biometrie und Epidemiologie, Institut für Soziale Medizin, Zentrum für Human- und Gesundheitswissenschaften, Freie Universität Berlin, Fabeckstr. 60-62, D-14195 Berlin; lungtung@zedat.fu-berlin.de

PROF. DR. ROLF R. ENGEL, Abteilung für Klinische Psychologie und Psychophysiologie, Klinik für Psychiatrie und Psychotherapie der LMU München, Nussbaumstr. 7, D-80336 München; re@psy.med.uni-muenchen.de

JEREMY FRANKLIN, Biometrie, Klinik I für Innere Medizin, Universität Köln, Herder Str. 52-54, D-50931 Köln; jeremy.franklin@Biometrie.uni-koeln.de

DR. MATTHIAS GREINER, International EpiLab, Danish Veterinary Institute, Bülowsvej 27, DK-1790 Copenhagen V; mgreiner@gmx.net

DR. HEIKO GROSSMANN, Psychologisches Institut IV, Westfälische Wilhelms-Universität, Fliednerstr. 21, D-48149 Münster; grossman@psy.uni-muenster.de

PROF. DR. JOACHIM HARTUNG, Fachbereich Statistik, Universität Dortmund, Vogelpothsweg 87, D-44221 Dortmund; hartung@statistik.uni-dortmund.de

PROF. DR. HEINZ HOLLING, Psychologisches Institut IV, Westfälische Wilhelms-Universität, Fliednerstr. 21, D-48149 Münster; holling@psy.uni-muenster.de

DR. ANDREAS JUETTING, Psychologisches Institut IV, Westfälische Wilhelms-Universität, Fliednerstr. 21, D-48149 Münster; juetting@psy.uni-muenster.de

DR. JÜRGEN KLEIN, Institut für Arbeits- und Sozialmedizin der Universität zu Köln, Joseph-Stelzmann-Str. 9, D-50924 Köln; juergen.klein@uni-koeln.de

DR. GUIDO KNAPP, Fachbereich Statistik, Universität Dortmund, Vogelpothsweg 87, D-44221 Dortmund; Knapp@statistik.uni-dortmund.de

DR. ARMIN KOCH, Bundesinstitut für Arzneimittel und Medizinprodukte, Seestr. 10, D-13353 Berlin; a.koch@bfarm.de

DR. KEPHER HENRY MAKAMBI, Fachbereich Statistik, Universität Dortmund, Vogelpothsweg 87, D-44221 Dortmund; makambi@amadeus.statistik.uni-dortmund.de

DR. UWE MALZAHN, AG Biometrie und Epidemiologie, Institut für Soziale Medizin, Zentrum für Human- und Gesundheitswissenschaften, Freie Universität Berlin, Fabeckstr. 60-62, D-14195 Berlin; malzahn@zedat.fu-berlin.de

PROF. DR. GEORG MATT, Department of Psychology, San Diego State University, San Diego, CA 92182-4611; gmatt@sciences.sdsu.edu

PROF. DR. JOACHIM ROEHMEL, Bundesinstitut für Arzneimittel und Medizinprodukte, Seestr. 10, D-13353 Berlin; j.roehmel@bfarm.de

DR. WILHELM SAUERBREI, Universitätsklinikum der Albert-Ludwigs-Universität Freiburg, Institut für Medizinische Biometrie und Medizinische Informatik, Stefan-Meier-Str. 26, D-79104 Freiburg; wfs@imbi.uni-freiburg.de

DR. PETER SCHLATTMANN, AG Biometrie und Epidemiologie, Institut für Soziale Medizin, Zentrum für Human- und Gesundheitswissenschaften, Freie Universität Berlin, Fabeckstr. 60-62, D-14195 Berlin; schlattmann@medizin.fu-berlin.de

DR. CLAUDIA SCHOECHLIN, Abteilung für Klinische Psychologie und Psychophysiologie, Klinik für Psychiatrie und Psychotherapie der LMU München, Nussbaumstr. 7, D-80336 München;
claudia.schoechlin@psy.med.uni-muenchen.de

DR. RALF SCHULZE, Psychologisches Institut IV, Westfälische Wilhelms-Universität, Fliednerstr. 21, D-48149 Münster; rs@psy.uni-muenster.de

PROF. DR. MARTIN SCHUMACHER, Universitätsklinikum der Albert-Ludwigs-Universität Freiburg, Institut für Medizinische Biometrie und Medizinische Informatik, Stefan-Meier-Str. 26, D-79104 Freiburg;
ms@imbi.uni-freiburg.de

GUIDO SCHWARZER, Universitätsklinikum der Albert-Ludwigs-Universität Freiburg, Institut für Medizinische Biometrie und Medizinische Informatik, Stefan-Meier-Str. 26, D-79104 Freiburg; sc@imbi.uni-freiburg.de

PROF. DR. KARL WEGSCHEIDER, Institut für Statistik und Ökonometrie, Universität Hamburg, Von-Melle-Park 5, D-20146 Hamburg;
wegsch@econ.uni-hamburg.de

DR. KLAUS WEIST, Institut für Hygiene und Umweltmedizin, Freie Universität Berlin, Hindenburgdamm 30, D-12203 Berlin; weist@ukbf.fu-berlin.de

PROF. DR. WERNER W. WITTMANN, Lehrstuhl Psychologie II, Universität Mannheim, Schloss, D-68131 Mannheim;
wittmann@tnt.psychologie.uni-mannheim.de

Subject Index

Author Index