

Ton J. Cleophas  
Aeilko H. Zwinderman

# Modern Meta-Analysis

Review and Update of Methodologies

# Modern Meta-Analysis

Ton J. Cleophas • Aeilko H. Zwinderman

# Modern Meta-Analysis

Review and Update of Methodologies



Springer

Ton J. Cleophas  
Albert Schweitzer Hospital  
Department of Medicine  
Sliedrecht, The Netherlands

Aeilko H. Zwinderman  
Department of Epidemiology and Biostatistics  
Academic Medical Center  
Amsterdam, Noord-Holland  
The Netherlands

ISBN 978-3-319-55894-3      ISBN 978-3-319-55895-0 (eBook)  
DOI 10.1007/978-3-319-55895-0

Library of Congress Control Number: 2017935817

© Springer International Publishing Switzerland 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature  
The registered company is Springer International Publishing AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Modern meta-analyses do more than combine the effect sizes of a series of similar studies. The term “meta” in meta-analysis can be interpreted as “beyond”, and meta-analyses are currently increasingly applied for any analysis beyond the primary analysis of studies. Terminologies like meta-learning, metacognition, meta-knowledge, higher order of thinking, and awareness of learning processes and thinking skills are used. We should add that nowadays, we have a big body of research data, thanks to the publication of one scholarly article every 20 seconds. Handling those big research data with powerful methods like meta-analytic forest plots, Bayesian networks, automatic data mining programs, etc. can now provide a more rapid learning process of essential issues and scientific progress. Very important and, even more so, with big data, the exchangeability assumption emphasized in the early 80th meta-analyses remains vital today: patient and study characteristics must be exchangeable and similar enough for studies to be compared.

This book was written for nonmathematical professionals of medical and health care, in the first place, but, in addition, for anyone involved in any field involving scientific research. Every methodology in this update will be explained with data examples, both hypothesized and real data. The authors have published many pretty innovative meta-analyses from the turn of the century till now. A list of international publications is given underneath. This edition will review the current state of the art and will use for that purpose the methodological aspects of these publications, in addition to other relevant methodological issues from literature. To readers requesting more background, theoretical, and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available: *Statistics Applied to Clinical Studies* 5th edition, 2012; *Machine Learning in Medicine: A Complete Overview*, 2015; *SPSS for Starters and 2nd Levelers* 2nd edition, 2015; *Clinical Data Analysis on a Pocket Calculator* 2nd edition, 2016; and *Understanding Clinical Data Analysis* from published research, 2016, all of which are edited by Springer Heidelberg, Germany.

Are there alternative works in the field? Yes, there are, particularly in the field of psychology. Psychologists have invented meta-analyses in 1970, and they have continuously written and updated methodologies ever since. Although very interesting, their work, just like the whole discipline of psychology, is rather explorative in nature, and so is the focus of their approach to meta-analysis. As such, they are not particularly involved in confirmatory placebo-controlled double-blind therapeutic clinical trials, and, despite their overwhelming productions and sometimes expensive software, they never address clinically important subjects like the meta-analysis with diagnostic tests, contrast coefficients, tetrachoric correlations, Bayesian networks, quasi-likelihood modeling, correlation coefficients to z transformations, confounding and interaction assessments, and other clinically relevant subjects. Then, there is the field of epidemiologists. Many of them are from the school of angry young men, who publish shocking news all the time, and JAMA and other publishers are happy to publish it. The reality is, of course, that things are usually not as bad as they seem. The recently published book entitled *Meta-analysis with R*, Springer Heidelberg, Germany, 2015, is lovely and, in addition, written by professional statisticians. A problem is that all analyses are with R software. R has a miserable menu program and requires lots of syntax to be learned. This is prohibitive to many clinical and other health professionals.

The current edition is a must-read textbook written by a very experienced mathematical statistician and an internist/clinical pharmacologist. It addresses new meta-analytical methodologies relevant to clinical research including diagnostic and therapeutic clinical trials and drug research. The book will consist, like our previous books, of many examples and step-by-step analyses using software like the free MetaXL from Excel, the SPSS work bench for automatic data mining entitled SPSS Modeler, the free Konstanz information miner (KNIME), and pocket calculator methods, if more convenient. In order for readers to perform their own analyses, SPSS data files are given in extras.springer.com.

## Published Meta-analyses from the Authors

1. Cleophas TJ, Niemeyer MG, Van der Wall EE. Wine, beer, and spirits and the risk of death and myocardial infarction. **Cardiologie** 1999; 6: 415–21.
2. Hornstra N, Hoogstraten B, Cleophas TJ, Van der Meulen J. Homocysteinemia and coronary artery disease, risk factor or not? A meta-analysis. **Am J Cardiol** 2000; 86: 1005–1009.
3. Cleophas TJ, Zwinderman AH. Beta-blockers and heart failure, a meta-analysis. **Int J Clin Pharmacol Ther** 2001; 39: 383–8.
4. Cleophas TJ, Zwinderman AH. Beta-blockers and heart failure, a meta-analysis. **Int J Clin Pharmacol Ther** 2001; 39: 562–3, letters.
5. Cleophas TJ, Grabowsky I, Niemeyer MG, Mäkel W. Nebivolol monotherapy in 3,147 patients with mild hypertension. **J Hypertens** 2001; 19: s2: 261.

6. Cleophas TJ, Van Marum R. Meta-analysis of efficacy and safety of second-generation dihydropyridine calcium channel blockers in chronic heart failure. *Am J Cardiol* 2001; 87: 487–90.
7. Cleophas TJ, Van Marum R, Cleophas AF, Zwinderman AH. Updated Meta-analysis of second generation dihydropyridine calcium channel blockers in chronic heart failure. *Br J Clin Pharmacol* 2001; abstract from FIGON Federatie Innovatief Geneesmiddelen Onderzoek, Lunteren, October 2001).
8. Cleophas TJ, Van Marum R, Cleophas AF, Zwinderman AH. Updated Meta-analysis of second generation dihydropyridine calcium channel blockers in chronic heart failure. *Cochrane Library* Heart Group, data online, 2001.
9. Cleophas TJ, Zwinderman AH. Primer in statistics. Meta-analysis. *Circulation* 2007; 115: 2870–5.
10. Masoor K, Cleophas TJ. Meta-analysis of heart failure in 39,505 patients with diabetes mellitus. *J Card Fail* 2009; 15: 305–9.
11. Atiqi R, Van Iersel C, Cleophas TJ. Accuracy of quantitative diagnostic tests. *Int J Clin Pharmacol Ther* 2009; 47: 153–9.
12. Atiqi R, Cleophas TJ, Van Bommel E, Zwinderman AH. Meta-analysis of recent studies on patients admitted to hospital due to adverse drug effects. *Int J Clin Pharmacol Ther* 2009; 47: 549 56.
13. Cleophas TJ, Zwinderman AH. Meta-analyses of diagnostic tests. *Clin Chem Lab Med* 2009; 47: 1351–4.
14. Cleophas TJ, Atiqi R, Zwinderman AH. Handling categories properly: a novel objective of clinical research. *Am J Ther* 2012; 19: 287–93.
15. Cleophas EP, Cleophas TJ. Multistage regression, a novel method for making better predictions from your efficacy data. *Am J Ther* 2011; Doi: 10.1097.
16. Sprangers S, Levin M, Cleophas T. Active recruitment for clinical trials (multinomial logistic regression), Abstract from FIGON Federatie Innovatie Geneesmiddelen Onderzoek, Lunteren, October 2006.
17. Akin S, Yetgin T, Brugts JJ, Dirkali A, Zijlstra F, Cleophas TJ. Effect of collaterals on deaths and re-infarctions in patients with coronary artery disease: a meta-analysis. *Neth Heart J* 2013; DOI 10.1007.
18. Van Bommel E, Cleophas TJ. Antihypertensive effect of potassium, a meta-analysis. *Int J Clin Pharmacol Ther* 2012; 50: 478–81.
19. Van Houwelingen HC, Zwinderman AH. A bivariate approach to meta-analysis. *Stat Med* 1993; 12: 2273–84.
20. Glas AS, Roos D, Deutkom M, Zwinderman AH, Bossuyt PM. Tumor markers in the diagnosis of primary bladder cancer: a systematic review. *J Urol* 2003; 169: 1975.
21. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; 58: 982–90.

Sliedrecht, The Netherlands  
Amsterdam, The Netherlands

Ton J. Cleophas  
Aeilko H. Zwinderman

# Contents

<b>1</b>	<b>Meta-analysis in a Nutshell . . . . .</b>	1
1.1	Introduction . . . . .	1
1.2	Objectives . . . . .	3
1.3	How to Perform a Meta-analysis? . . . . .	5
1.4	Scientific Rigor, Rule 1 . . . . .	8
1.5	Scientific Rigor, Rule 2 . . . . .	10
1.6	Scientific Rigor, Rule 3 . . . . .	11
1.7	Scientific Rigor, Rule 4 . . . . .	12
1.8	First Pitfall . . . . .	15
1.9	Second Pitfall . . . . .	16
1.10	Third Pitfall . . . . .	19
1.11	Benefits and Criticisms of Meta-analyses . . . . .	19
1.12	Conclusion . . . . .	21
	Reference . . . . .	22
<b>2</b>	<b>Mathematical Framework . . . . .</b>	23
2.1	Introduction . . . . .	23
2.2	General Framework . . . . .	24
2.3	Continuous Outcome Data, Mean and Standard Deviation . . . . .	25
2.3.1	Means and Standard Deviation (SD) . . . . .	25
2.4	Continuous Outcome Data, Strictly Standardized Mean Difference (SSMD) . . . . .	26
2.5	Continuous Outcome Data, Regression Coefficient and Standard Error . . . . .	27
2.6	Continuous Outcome Data, Student's T-Value . . . . .	28
2.7	Continuous Outcome Data, Correlation Coefficient ( $R$ or $r$ ) and Its Standard Error . . . . .	29
2.8	Continuous Outcome Data, Coefficient of Determination $R^2$ or $r^2$ and Its Standard Error . . . . .	31
2.9	Binary Outcome Data, Risk Difference . . . . .	32

2.10	Binary Outcome Data, Relative Risk . . . . .	32
2.11	Binary Outcome Data, Odds Ratio . . . . .	33
2.12	Binary Outcome Data, Survival Data . . . . .	33
2.13	Pitfalls, Publication Bias . . . . .	34
2.14	Pitfalls, Heterogeneity . . . . .	35
2.15	Pitfalls, Lack of Sensitivity . . . . .	38
2.16	New Developments . . . . .	39
2.17	Conclusions . . . . .	40
	Reference . . . . .	41
<b>3</b>	<b>Meta-analysis and the Scientific Method . . . . .</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	Example 1, the Potassium Meta-analysis of the Chap. 6 . . . . .	44
3.3	Example 2, the Calcium Channel Blocker Meta-analysis of the Chap. 6 . . . . .	45
3.4	Example 3, the Large Randomized Trials Meta-analyses of the Chap. 6 . . . . .	45
3.5	Example 4, the Diabetes and Heart Failure Meta-analysis of the Chap. 7 . . . . .	46
3.6	Example 5, the Adverse Drug Effect Admissions and the Type of Research Group Meta-analysis of the Chap. 8 . . . . .	46
3.7	Example 6, the Coronary Events and Collaterals Meta-analysis of the Chap. 8 . . . . .	47
3.8	Example 7, the Diagnostic Meta-analysis of Metastatic Lymph Node Imaging of the Chap. 10 . . . . .	47
3.9	Example 8, the Homocysteine and Cardiac Risk Meta-analysis of the Chap. 11 . . . . .	48
3.10	Conclusions . . . . .	48
	References . . . . .	49
<b>4</b>	<b>Meta-analysis and Random Effect Analysis . . . . .</b>	<b>51</b>
4.1	Introduction . . . . .	51
4.2	Visualizing Heterogeneity . . . . .	53
4.3	Binary Outcome Data, Fixed Effect Analysis . . . . .	55
4.4	Binary Outcome Data, Random Effect Analysis . . . . .	56
4.5	Continuous Outcome Data, Fixed Effect Analysis . . . . .	58
4.6	Continuous Outcome Data, Random Effect Analysis . . . . .	60
4.7	Conclusions . . . . .	61
	Reference . . . . .	62
<b>5</b>	<b>Meta-analysis with Statistical Software . . . . .</b>	<b>63</b>
5.1	Introduction . . . . .	63
5.2	Using Online Meta-analysis Calculators and MetaXL Free Meta-analysis Software . . . . .	63
5.3	Continuous Outcome Data, Online Meta-analysis Calculator . . . . .	64

5.4	Binary Outcome Data, MetaXL Free Meta-analysis	
	Software . . . . .	67
5.4.1	Traditional Random Effect Analysis . . . . .	68
5.4.2	Quasi Likelihood (Invert Variance Heterogeneity (IVhet)) Modeling for Heterogeneity . . . . .	72
5.5	Conclusion . . . . .	75
	Reference . . . . .	77
<b>6</b>	<b>Meta-analyses of Randomized Controlled Trials . . . . .</b>	<b>79</b>
6.1	Introduction . . . . .	79
6.2	Example 1: Single Outcomes . . . . .	81
6.3	Example 1, Confirming the Scientific Question . . . . .	85
6.4	Example 2: Multiple Outcomes . . . . .	85
6.5	Example 2, Handling Multiple Outcomes . . . . .	87
6.6	Example 3, Large Meta-analyses Without Need for Pitfall Assessment . . . . .	88
6.7	Conclusion . . . . .	90
	Reference . . . . .	91
<b>7</b>	<b>Meta-analysis of Observational Plus Randomized Studies . . . . .</b>	<b>93</b>
7.1	Introduction and Example . . . . .	93
7.2	Sound Clinical Arguments and Scientific Question . . . . .	94
7.3	Summary Statistics . . . . .	95
7.4	Pooled Results . . . . .	96
7.5	Heterogeneity Assessments . . . . .	98
7.6	Publication Bias Assessments . . . . .	98
7.7	Robustness Assessments . . . . .	99
7.8	Improved Information from the Combined Meta-analysis . . . . .	99
7.9	Conclusion . . . . .	100
	Reference . . . . .	100
<b>8</b>	<b>Meta-analysis of Observational Studies . . . . .</b>	<b>101</b>
8.1	Introduction . . . . .	101
8.2	Prospective Open Evaluation Studies . . . . .	102
8.3	Example 1, Event Analysis in Patients with Collateral Coronary Arteries . . . . .	103
8.4	Example 1, the Scientific Method . . . . .	103
8.5	Example 1, Publication Bias . . . . .	104
8.6	Example 1, Pooled Results, Tests for Heterogeneity and Robustness . . . . .	104
8.7	Example 1, Meta-regression Analysis . . . . .	106
8.8	Conclusions . . . . .	108
8.9	Example 2, Event Analysis of Iatrogenic Hospital Admissions . . . . .	108
8.10	Example 2, the Scientific Method . . . . .	108
8.11	Example 2, Publication Bias . . . . .	109

8.12	Example 2, Overall Results and Heterogeneity and Lack of Robustness . . . . .	110
8.13	Example 2, Meta-regression . . . . .	112
8.14	Example 2, Conclusions . . . . .	113
8.15	Conclusion . . . . .	113
	Reference . . . . .	114
<b>9</b>	<b>Meta-regression . . . . .</b>	<b>115</b>
9.1	Introduction . . . . .	115
9.2	Example 1, Continuous Outcome . . . . .	115
9.2.1	Exploratory Purpose . . . . .	116
9.2.2	Confounding . . . . .	117
9.2.3	Interaction . . . . .	119
9.3	Example 2, Binary Outcome . . . . .	121
9.3.1	Exploratory Purpose . . . . .	121
9.3.2	Confounding . . . . .	122
9.3.3	Interaction . . . . .	123
9.4	Conclusion . . . . .	124
	Reference . . . . .	126
<b>10</b>	<b>Meta-analysis of Diagnostic Studies . . . . .</b>	<b>127</b>
10.1	Introduction . . . . .	127
10.2	Diagnostic Odds Ratios (DORs) . . . . .	129
10.3	Example . . . . .	129
10.4	Constructing Summary ROC Curves . . . . .	132
10.5	Alternative Methods . . . . .	133
10.6	Conclusions . . . . .	133
	References . . . . .	134
<b>11</b>	<b>Meta-Meta-analysis . . . . .</b>	<b>135</b>
11.1	Introduction . . . . .	135
11.2	Example 1, Meta-Meta-analysis for Re-assessment of the Pitfalls of the Original Meta-analyses . . . . .	136
11.3	Example 2, Meta-Meta-analysis for Meta-learning Purposes . . . . .	141
11.4	Conclusion . . . . .	142
	Reference . . . . .	143
<b>12</b>	<b>Network Meta-analysis . . . . .</b>	<b>145</b>
12.1	Introduction . . . . .	145
12.2	Example 1, Lazarou-1 (JAMA 1998; 279: 1200–5) . . . . .	146
12.3	Example 2, Atiqi (Int J Clin Pharmacol Ther 2009; 47: 549–56) . . . . .	148
12.4	Example 3, Lazarou-1 and Atiqi (JAMA 1998; 279: 1200–5, and Int J Clin Pharmacol Ther 2009; 47: 549–56) . . . . .	151
12.5	Example 4, Lazarou-1 and -2 (JAMA 1998; 279: 1200–5) . . . . .	152
12.6	Conclusion . . . . .	155
	Reference . . . . .	155

<b>13 Random Intercepts Meta-analysis . . . . .</b>	157
13.1 Introduction . . . . .	157
13.2 Example, Meta-analysis of Three Studies . . . . .	159
13.3 Conclusion . . . . .	165
Reference . . . . .	165
<b>14 Probit Meta-regression . . . . .</b>	167
14.1 Introduction . . . . .	167
14.2 Example . . . . .	167
14.3 Conclusion . . . . .	175
Reference . . . . .	175
<b>15 Meta-analysis with General Loglinear Models . . . . .</b>	177
15.1 Introduction . . . . .	177
15.2 Example, Weighted Multiple Linear Regression . . . . .	177
15.3 Example, General Loglinear Modeling . . . . .	180
15.4 Conclusion . . . . .	183
Reference . . . . .	183
<b>16 Meta-analysis with Variance Components . . . . .</b>	185
16.1 Introduction . . . . .	185
16.2 Example 1 . . . . .	185
16.3 Example 2 . . . . .	187
16.4 Example 3 . . . . .	189
16.5 Conclusion . . . . .	192
Reference . . . . .	193
<b>17 Ensembled Correlation Coefficients . . . . .</b>	195
17.1 Introduction . . . . .	195
17.2 Ensemble Learning with SPSS Modeler . . . . .	196
17.3 Example . . . . .	197
17.4 Conclusion . . . . .	203
Reference . . . . .	204
<b>18 Ensembled Accuracies . . . . .</b>	205
18.1 Introduction . . . . .	205
18.2 Ensemble Learning with SPSS Modeler . . . . .	206
18.3 Example . . . . .	207
18.4 Conclusion . . . . .	215
Reference . . . . .	216
<b>19 Multivariate Meta-analysis . . . . .</b>	217
19.1 Introduction . . . . .	217
19.2 Example 1 . . . . .	218
19.3 Example 2 . . . . .	222
19.4 Example 3 . . . . .	226
19.5 Conclusions . . . . .	231
Reference . . . . .	231

<b>20</b>	<b>Transforming Odds Ratios into Correlation Coefficients . . . . .</b>	233
20.1	Introduction . . . . .	233
20.2	Unweighted Odds Ratios as Effect Size Calculators . . . . .	234
20.3	Regression Coefficients and Correlation Coefficients as Replacement of Odds Ratios . . . . .	234
20.4	Examples of Approximation Methods for Computing Correlation Coefficients from Odds Ratios . . . . .	238
20.4.1	The Yule Approximation . . . . .	238
20.4.2	The Ulrich Approximation . . . . .	239
20.5	Computing Tetrachoric Correlation Coefficients from an Odds Ratio . . . . .	239
20.6	Conclusion . . . . .	241
	Reference . . . . .	242
<b>21</b>	<b>Meta-analyses with Direct and Indirect Comparisons . . . . .</b>	243
21.1	Introduction . . . . .	243
21.2	Challenging the Exchangeability Assumption . . . . .	244
21.3	Frequentists' Methods for Indirect Comparisons . . . . .	244
21.4	The Confidence Intervals Methods for Indirect Comparisons . . . . .	245
21.5	Real Data Examples . . . . .	247
21.6	Conclusion . . . . .	248
	Reference . . . . .	248
<b>22</b>	<b>Contrast Coefficients Meta-analysis . . . . .</b>	249
22.1	Introduction . . . . .	249
22.2	Example . . . . .	250
22.3	Fixed Effect Meta-analysis . . . . .	252
22.4	Random Effect Meta-analysis . . . . .	253
22.5	Principles of Linear Contrast Testing . . . . .	254
22.6	Null Hypothesis Testing of Linear Contrasts . . . . .	255
22.7	Pocket Calculator One-Way Analysis of Variance (ANOVA) of Linear Contrasts . . . . .	256
22.8	Contrast Testing Using One Way Analysis of Variance on SPSS Statistical Software . . . . .	257
22.9	Conclusion . . . . .	258
	References . . . . .	259
<b>23</b>	<b>Meta-analysis with Evolutionary Operations (EVOPs) . . . . .</b>	261
23.1	Introduction . . . . .	261
23.2	Example of the Meta-analysis of Three Studies Assessing Determinants of Infectious Disease . . . . .	262
23.3	First Study . . . . .	263
23.4	Second Study . . . . .	263
23.5	Third Study . . . . .	265
23.6	Meta-analysis of the Above Three Studies . . . . .	266
23.7	Conclusion . . . . .	267
	Reference . . . . .	267

<b>24</b>	<b>Meta-analysis with Heterogeneity as Null-Hypothesis . . . . .</b>	269
24.1	Introduction . . . . .	269
24.2	Heterogeneity Rather Than Homogeneity of Studies as Null-Hypothesis . . . . .	270
24.3	Frequency Distribution of the Treatments . . . . .	272
24.4	Beneficial Effects of the Treatments, Histograms . . . . .	272
24.5	Beneficial Effects of Treatments, Clusters . . . . .	273
24.6	Beneficial Effects of Treatments, Network of Causal Associations Displayed as a Web . . . . .	274
24.7	Beneficial Effects of Treatments, Decision Trees . . . . .	275
24.8	Beneficial Effects of Treatments, Accuracy Assessment of Decision Tree Output . . . . .	276
24.9	Conclusions . . . . .	276
	Reference . . . . .	277
<b>25</b>	<b>Meta-analytic Thinking and Other Spin-Offs of Meta-analysis . . . . .</b>	279
25.1	Introduction . . . . .	279
25.2	Meta-learning . . . . .	279
25.3	Meta-analytic Graphing . . . . .	280
25.4	Meta-analytic Thinking in Writing Study Protocols and Reports . . . . .	282
25.5	Meta-analytic Forest Plots of Baseline Patient Characteristics . . . . .	284
25.6	Meta-analysis of Forest Plots of Propensity Scores . . . . .	287
25.7	Meta-analytic Thinking: Effect Size Assessments of Important Scientific Issues Other Than the Main Study Outcomes . . . . .	290
25.8	Forest Plots for Assessing and Adjusting Baseline Characteristic Imbalance . . . . .	292
25.9	Sensitivity Analysis . . . . .	293
25.10	Pooled Odds Ratios for Multidimensional Outcome Effects . . . . .	294
25.11	Ratios of Odds Ratios for Subgroup Analyses . . . . .	296
25.12	Conclusion . . . . .	298
	Reference . . . . .	298
<b>26</b>	<b>Novel Developments . . . . .</b>	299
26.1	Introduction, Condensed Review of the Past . . . . .	299
26.2	Condensed Review of the Current Edition and Novel Developments . . . . .	300
26.3	Meta-analysis of Studies Tested with Analyses of Variance . . . . .	301
26.4	Meta-analyses of Crossover Trials with Binary Outcomes . . . . .	303
26.5	Equivalence Study Meta-analysis . . . . .	304

26.6	Agenda-Driven Meta-analyses . . . . .	306
26.7	Hills' Plurality of Reasoning Statement, Evidence Based Medicine Avant la Lettre . . . . .	307
26.8	Conclusion . . . . .	307
	Reference . . . . .	308
<b>Index</b> . . . . .		<b>309</b>

# **Chapter 1**

## **Meta-analysis in a Nutshell**

### **Meta-analyses, the Pinnacle of Science or an Error-Ridden Exercise**

**Abstract** A meta-analysis is a systematic review with pooled outcome data. The current chapter gives a summary of methods for the purpose. Four scientific rules and three pitfalls are reviewed. The main benefit of meta-analysis is that it provides a pooled outcome with increased precision. The main criticisms are that so far they were not good at predicting subsequent large trials, and at predicting serious adverse effects of medicines.

#### **1.1 Introduction**

A single study is no more than tentative evidence, and has to be confirmed by independent additional studies. Narrative reviews of studies from the literature suffers from subjectivity, and meta-analysis methodology requires a number of pretty objective criteria. In addition, the use of effect sizes and 95% confidence intervals instead of p-values, has a lot of advantages. For example, p-values are pretty hard to non-statisticians, and they are commonly misunderstood. They are based on probabilities, and not just on ordinary probabilities but conditional probabilities. A p-value  $>0.05$  does not mean that the null-hypothesis is true. It is the chance of making a type error of finding a difference from zero where there is none. A p-value actually means: if and only if no difference from a zero effect is in the data, then you will have a probability of  $<5\%$  to find a p-value of 5% or less. A mean effect size and 95% confidence are much easier to understand. For example, they tell you what effect you can expect on average and between what intervals your effects in the future will be 95% of the times. There is of course the disadvantage that currently research, education, and software universally emphasize p-values, but this is just a matter of time. The effect size approach is a core characteristic of meta-analysis and will be addressed in most chapters of this edition. This edition was written for physicians and other nonmathematical health workers. A condensed description of methodologies is given. Meta-analyses:

- are secondary, otherwise called post/hoc, analyses,
- test, however, primary, not secondary, hypotheses,
- use probability statements, that are probably more valid than those of other secondary studies are.

A meta-analysis is a systematic review of trials with pooled data. How long do they exist? In the early 70s psychologists were the first to perform systematic reviews, without pooled data. Only pretty recently, since 1995, more homogeneous trials were published, and more pooling could take place.

Why do we perform meta-analyses? This is pretty obvious: with more data, we are likely to obtain more certainty. However, also more differences may be observed. The essentials of meta-analysis need just a few words. All that is needed, is the rules of scientific rigor and a brief list of pitfalls is given. Scientific rigor requires that we stick to

- (1) a clearly defined prior hypothesis,
- (2) a thorough search of trials,
- (3) strict inclusion criteria,
- (4) uniform guidelines for data analysis.

A brief list of pitfalls includes

- (1) checking publication bias,
- (2) checking heterogeneity,
- (3) checking robustness.

The consequence of pooling the information from multiple studies into a single report carries a lot of general risks in addition to the above traditional pitfalls.

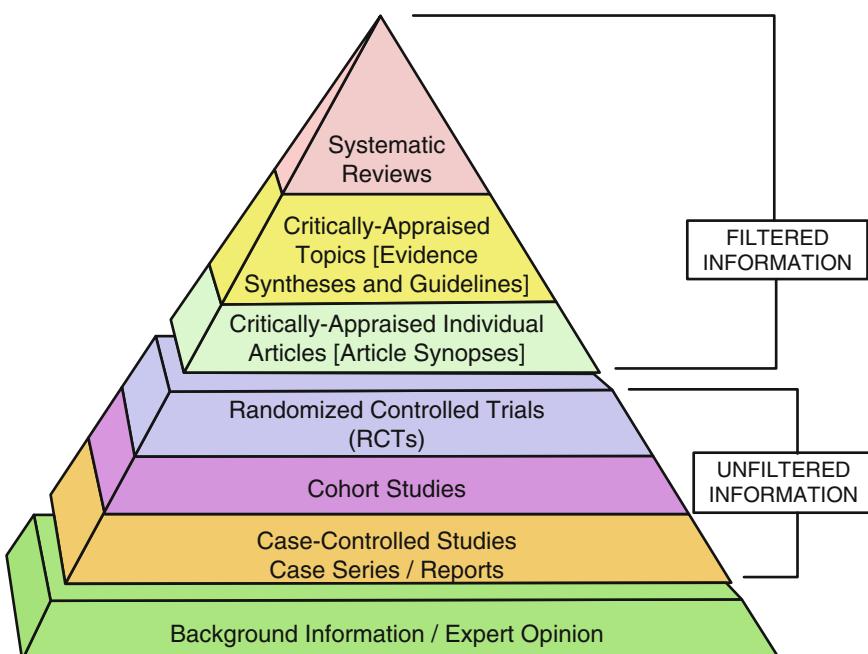
1. Integrated results means that from the original studies a lot of information is removed. At its extremes, large amounts of varied information is reduced to a single value per study and then it is concluded, that every relevant detail from the individual studies is covered. This can, of course, never be true. It can only be true in so far as the main endpoint of the studies is considered.
2. Publication bias is a problem, but just the tip of the iceberg. The identification and selection of relevant studies is another problem. Not all search machines are the same, and small differences carry a lot of weight in the final search result (Walker et al, Meta-analysis its strengths and weaknesses, Cleveland Clinic Journal of Medicine 2008; 75: 431–9).
3. The background information of original studies including patient recruitment procedures, type of blinding, all of the inclusion and exclusion criteria of the original studies can never be fully taken into account in a meta-analysis.
4. Current clinical trial journals require manuscripts with limited text and tables, and a lack of information is the consequence. The summary results as given, will be a limitation to meta-analyses, if they aim at reporting subgroup effects from original studies.

We should add, that, sometimes, the meta-analysis is there to produce the qualitative final endpoint of multiple quantitative studies. And so, the meta-analysis results are to tell us, whether or not all of the simultaneous effects of a host of tentative research does work or does not do so.

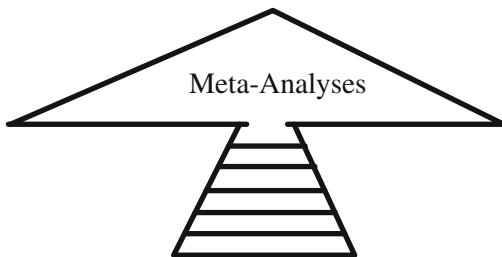
## 1.2 Objectives

In 1982 thrombolytic therapy for acute coronary syndromes was controversial. In a meta-analysis of 7 clinical trials Stampfer et al. found a reduced risk of mortality of 0.80 with a 95% confidence interval of 0.68–0.95 (Stampfer et al, N Engl J Med 1982; 307: 1180–2). Because of unfamiliarity of the cardiological community with the methodology of meta-analysis, these findings were not accepted until 1986, when a large clinical trial confirmed the conclusions of this meta-analysis (GISSI Study Group, Lancet 1986; I: 397–402), and streptokinase became widely applied.

Nowadays, a large body of clinical research makes meta-analysis feasible, and meta-analysis is widely recognized as part of evidence based medicine. Some even name it the top of the evidence based medicine pyramid (Tebala GD, What is the future of biomedical research, in: Medical Hypotheses, Pub Med 2015; ID 26194725). So much so, that we may also question: are they used too often, and what is their added benefit, if everything has already been said. And, are they sometimes used for reinforcing curricula vitae, rather than adding scientific knowledge. Is the pyramid of evidence based medicine top heavy, because it is based on an amount of evidence that little grows. The pinnacle of evidence based medicine (Cochrane) starts being liable to collapse as shown underneath. Is the meta-analysis the platinum standard of evidence as expressed by Stegenga. Or, do we have to argue that meta-analysis falls far short of that standard (Stegenga, Studies in history and philosophy of biological and biomedical sciences 2011; doi: 10.1016).



Underneath a graph after the Cochrane pyramid of evidence-based medicine (copied from the public domain) is given.



We will start making some statements that are pretty much consensus nowadays.  
Meta/analyses:

- are secondary, otherwise called post/hoc, analyses,
- test, however, primary, not secondary, hypotheses,
- have probability statements that may, therefore, be more valid than those of secondary studies.

Not all of the medical journals publish meta-analyses, and those who do, do not do so uniformly.

- The J Am Med Assoc, Br Med J, and Lancet in 2000 published 14, 24, and 7 meta-analyses (all of them performed according to the famous Cochrane guidelines ([training.cochrane.org/handbook](http://training.cochrane.org/handbook))).
- The N Engl J Med had no meta-analyses in its index (61 review monographies).
- Specialists' journals including the J Am Coll Cardiol, Diabet Care, J Oncol, Gastroenterology, Angiol, J Neurol & Neurosurg Psychiatr, J Am Geriat Soc, J Clin Endocrin Metab, 8 journals all-in published 1–3 meta-analyses each, and they were not according to the above guidelines, because publication bias was tested 0 times, heterogeneity 2 times, and sensitivity 2 times.

Yet, pretty much agreement exists on the points given underneath:

- the method is increasingly appreciated by the medical community,
- used by regulatory bodies,
- reduces the boundaries of uncertainty,
- and the development of new drugs benefits from it.

This chapter has been written for physicians and other health workers, non-mathematicians, and, for their benefit, the authors will refrain from using mathematical equations. The basic principles will be given particular attention:

- publication bias meaning that negative trials are generally not published, and,
- heterogeneity between studies selected for the meta-analysis is due to different methods and different populations applied.

It is the authors' impression as professors in medical statistics for over 30 years, that among professionals a lot of misunderstanding exists regarding meta-analysis. Two anecdotes will be given.

- Anecdote 1

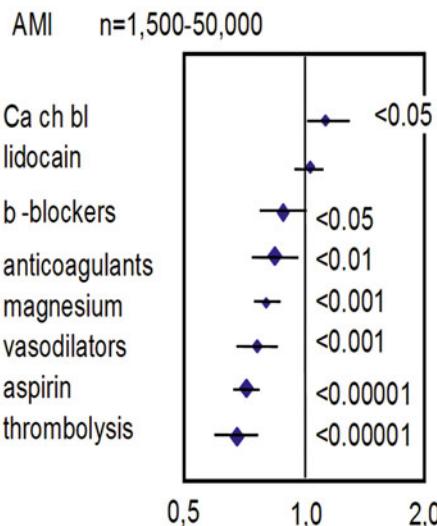
Groningen (Netherlands) pharmacologists performed a meta-analysis of efficacy of ACE-inhibitors, and said they excluded publication bias by thoroughly searching Medline, and heterogeneity by strict inclusion criteria.

- Anecdote 2

Professor vd Broucke from Leyden Netherlands published the opinion in 2000 in Neth J Med that negative trials are irrelevant; The founders of meta-analysis Oxmann, Guyatt, Chalmers from UK (letters 1996; 40: 509–510 Neth J Med) responded furiously that negative trials were to complete meta-analyses.

What is a meta-analysis? It is a systematic review of trials with pooled data. How long do they exist? In the early 70s psychologists were the first to perform systematic reviews, without pooled data. Only pretty recently, since 1995, more homogeneous trials were published, and more pooling could take place. Why do we perform meta-analyses? This is pretty obvious: with more data, we are likely to obtain more certainty. However, also more differences between subgroups will be observed.

### 1.3 How to Perform a Meta-analysis?



The above figure is not a meta-analysis, but it is an overview of the chief results of 8 large meta-analyses. The pooled data of the large meta-analyses of the treatment myocardial infarct (AMI) are on the x-axis, and they are expressed as odds ratio (OR) ( $\pm 95\%$  confidence intervals). The odds ratio = the chance of mortality from MI in users of the medicines/chance of mortality in non-users. If the OR  $< 1$ , then the medicine will be efficacious, if  $> 1$ , the it will not be efficacious (the size of black squares correspond to sizes of samples included in the meta-analyses). Acute PTCA (angioplasty) was not added, but result would of course have probably been even better.

The term odds is hard and often misunderstood. Nobody except gamblers immediately understand the meaning of odds. The term stems from gambling, either you win or you lose, there is nothing in between. The real meaning of the ratio of two odds are even harder, but they are usually interpreted as the ratio of two chances: the chance of event with active treatment/chance of event with placebo. The problem with true chances is, that they run from zero to one, while odds run from zero to infinite. Statistical software programs have difficulties with chances and work fine with odds. The use of odds ratios has expanded through their use as major effect size estimators in meta-analyses. Odds may be hard to understand, but we will use it here to indicate the likelihood = chance = probability = risk that an event will occur divided by the chance that it won't. A contingency table, otherwise called  $2 \times 2$  interaction table, is helpful.

<u>Contingency table</u>	numbers subjects who died	numbers subjects who did not die
Test treatment (group 1)	a	b
Control treatment (group 2)	c	d

The proportion of subjects who died in group 1 (or the risk (R) or probability of having an effect) equals

$$p = a/(a + b),$$

in group 2

$$p = c/(c + d),$$

The ratio of  $a/(a + b)$  and  $c/(c + d)$  is called the risk ratio or relative risk (RR). An approach different from the risk ratio approach is the odds approach, where

$$\begin{matrix} a/b \\ c/d \end{matrix}$$

are the odds and their ratio is the odds ratio (OR).

In clinical trials we use ORs as surrogates of the RRs, because here  $a/a + b$  are nonsense. An example will be given.

	treatment	control	group	whole population
Sleepiness(n)	32	a	4	b
No sleepiness(n)	24	c	52	d

We assume, that the control group is just a sample from population, but ratio

$$b/d$$

is that of the population. So, suppose

$$4 = 4000 \text{ and}$$

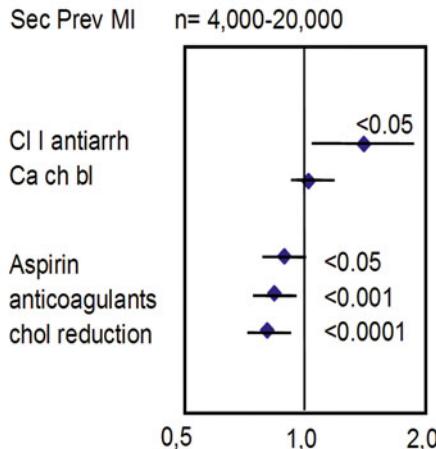
$$52 = 52000, \text{ then}$$

$$\frac{a/a + b}{c/c + d} \text{ is close to}$$

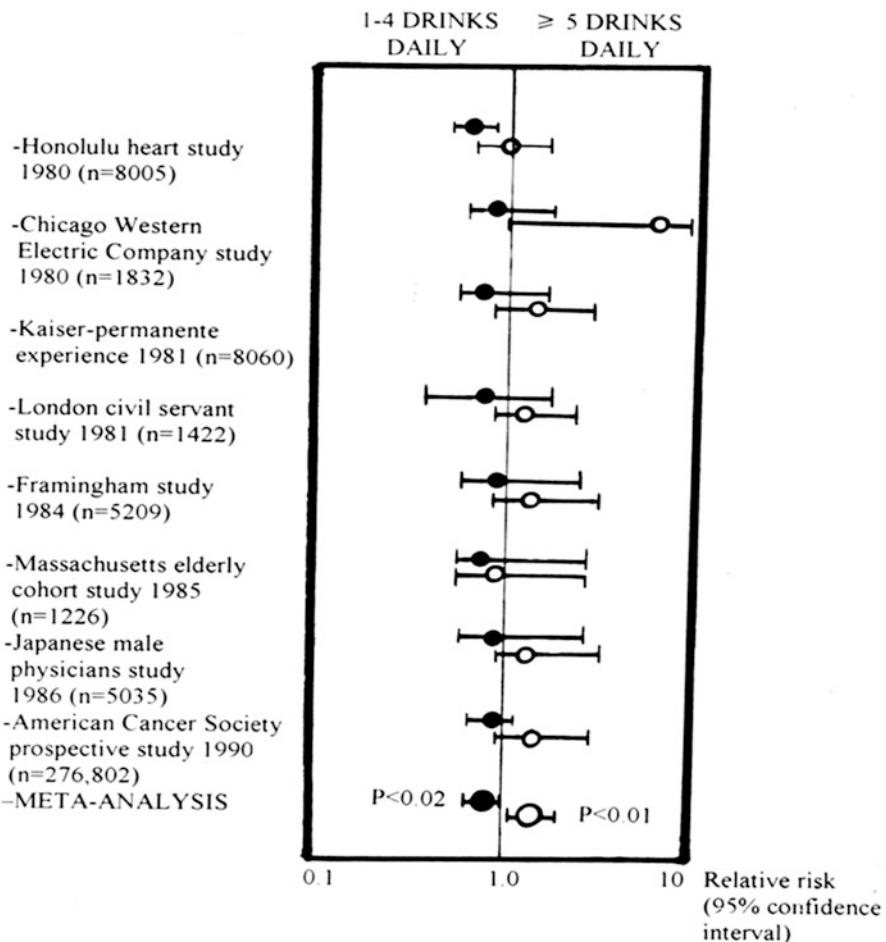
$$\frac{a/b}{c/d} = \text{the RR of the population}$$

Underneath is another example of many meta-analyses giving the pooled data of large meta-analyses for secondary prevention of myocardial infarction (MI).

On the x-axis we have the odds ratio (OR) = chance second MI among users/ chance second MI among non-users of the medicines given. An OR <1 indicates that the compound is efficacious, >1, not efficacious.



Not only controlled clinical trials, but also epidemiological studies can be meta-analyzed. The underneath figure gives an example. It shows:



retrospective cohort studies of risk factor carriers versus no-risk factor carriers (alcohol intake as a risk factor for coronary heart disease). On the x-axis the relative risk (RR) instead of odds ratio (OR) is given here. The RR is the chance of MI in drinkers/chance of MI in non-drinkers. Two risk factors are assessed: the risk of MI with 1–4 daily drinks versus no drinks (closed circles), and the risk of MI with  $\geq 5$  daily drinks (open circles). A RR  $< 1$  means that the risk factor protects, a RR  $> 1$  means that the risk factor does not protect. One to four daily drinks significantly protects from MI, over 5 daily drinks significantly increases the risk of MI.

## 1.4 Scientific Rigor, Rule 1

The scientific method will be given full attention in the Chap. 3. Briefly, it can be defined as “reformulating your scientific question into a hypothesis and testing it against control”. Here the subject of “scientific rigor” will be reviewed. This is a

subject as important as the scientific method, but it is different from the scientific method, and it means “scientific stiffness”: any scientific research will only be valuable, if its methodologies are as stiff as a dead body. Generally, important matters need few words. The underneath photocopy of Michelangelo’s Moses in the San Pietro in Vincoli Rome Italy shows the tables of the ten commandments, with a text of only 279 words. The American Declaration of Independence is another document of few words, only 300. However, current documents like European Community (EC) Directives regarding the import of caramel consists of many thousands of words (26,000 words). Many books have been written about guidelines regarding meta-analyses. However, the essentials of meta-analyses need few words. All that is needed, is the rules of the scientific rigor, and a brief list of pitfalls.



Scientific rigor means stiffness of the methodological approach to scientific research. Scientific processes are as stiff as a dead body with rigor mortis. Scientific rigor requires that we stick to

- (1) a clearly defined prior hypothesis,
- (2) to a thorough search of trials,
- (3) to strict inclusion criteria for trials, and
- (4) to uniform guidelines for data analysis.

The first rule is a clearly defined prior hypothesis. Why prior, why not posterior hypothesis? Good research starts with a prior hypothesis. This is tested prospectively at a probability of 0.05 (5% chance it is untrue). The problem with posterior hypotheses is, that one is easily seduced to test posterior hypotheses more often,

which increases the chance of success. For example testing 20 times or so. This practice is called data dredging. Statistical significances are then found by chance. It is like gambling, gambling 20 times at 5% chance gives you up to 70% chance of success instead of 5%.

## 1.5 Scientific Rigor, Rule 2



The second rule is a thorough search of the trials. A systematic procedure is required for that purpose. A checklist is helpful, just like the checklists used by aircraft personnel before take off. As shown in the above figure, without proper checking the checklists things may severely go wrong.

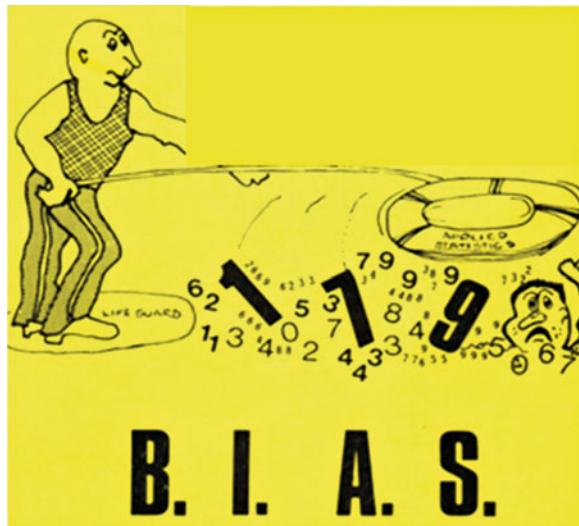
### Checklist for Medline Search

step	search term	step	search term
<b>indication</b>		<b>type</b>	<b>publication</b>
1	tendinitis.sh.	22	randomized controlled trial.pt.
2	elbow.s.	23	controlled clinical trial.pt.
3	elbow joint.sh.	24	randomized controlled trials.sh.
4	2 or 3	25	random allocation.sh.
5	1 and 4	26	double blind method.sh.
6	tennis elbow.sh.	27	single blind method.sh.
7	5 or 6	28	22 or 23 or 24 or 25 or 26 or 27
8	epicondylitis.tw.	29	(animal no (human and animal)).sh.
9	elbow.tw.	30	28 not 29
10	7 or 8 or 9	31	clinical trial.pt.
		32	exp clinical trials.sh.
<b>intervention</b>		33	((clinS adj25 trial\$).tw.
11	injections.sh.	34	((singlS or doublS or triplS) adj25 (blind\$ or mask\$).tw.
12	inject\$.tw.	35	placebos.sh.
13	infiltr\$.tw.	36	placeboS.tw.
14	exp glucocorticosteroids.sh.	37	randomS.tw.
15	triamcinolon\$.tw.	38	research design.sh.
16	hydrocortison\$.tw.	39	volunteerS.tw.
17	methylprednisolon\$.tw.	40	31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39
18	betamethason\$.tw.	41	40 not 29
19	lidocain\$.tw.	42	41 not 30
20	bupivacain\$.tw.	43	30 or 42
21	11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20		
			<i>indicatie, interventie en onderzoekstype</i>
		44	10 and 21 and 43

However, with a checklist you will be on right track in no time, but as shown above, it will still take 45 steps to take. Start logging in the chief indication of your search (could be a diagnosis group). Then log in “intervention” (could be a medicine)

Then take the appropriate steps to arrive at randomized controlled trials (RCTs).

## 1.6 Scientific Rigor, Rule 3



Strict inclusion criteria are needed. It reduces the chance of biases.

Bias means a systematic error that no one is aware of. Less bias will be in the research, if it is blinded, has proper statistics, has proper ethics, and proper descriptions of the methods of (re)search.

## 1.7 Scientific Rigor, Rule 4



Uniform guidelines for data analysis are indispensable. We are talking of statistics. Professor Bradford Hills from London UK once said: investigators use statistics, as a drunk uses a lamppost, for support rather than illumination. Appropriate statistics is, however, a powerful aid to prevent erroneous conclusions. It should not be too complicated. It should not be “dredging for statistical significances”. It should test prior, rather than posterior, hypotheses.

A study of continuous outcome data produces a sample mean, or rather “mean difference”, since controlled studies generally produce two means. For statistical testing an unpaired t-test of the sum of all “mean differences” is adequate.

$$t = \frac{\text{mean difference}_1 + \text{mean difference}_2 + \text{mean difference}_3 + \dots}{\sqrt{\text{SEM}_1^2 + \text{SEM}_2^2 + \text{SEM}_3^2 + \dots}}$$

with degrees of freedom =  $n_1 + n_2 + n_3 + \dots + n_k - k$ ,  $n_i$  = sample size ith sample,  $k$  = number of samples, SEM = standard error of the mean difference. If the standard deviations are very different in size, e.g., if one is twice the other, then a more adequate calculation of the pooled standard error is as follows. This formula gives greater weight to the pooled SEM the greater the samples. Similarly, if the

samples are very different in size, then a more adequate calculation of the nominator of  $t$  is as follows.

$$k \left( \frac{\text{mean difference}_1 n_1 + \text{mean difference}_2 n_2 + \dots}{n_1 + n_2 + \dots} \right)$$

Instead of a multiple samples  $t$ -test, a Cochran Q-test with chi-square statistics is possible (see Chap. 2). Probably, 99% of meta-analyses make use of proportions rather than continuous data, even if original studies provided predominantly the latter particularly for efficacy data (mean fall in blood pressure etc.). This is so both for efficacy and safety meta-analyses. Sometimes data have to be remodeled from quantitative into binary ones for that purpose. The *weighted average effect* is used as assessment of overall effect and for testing the significance of difference from zero (the null-hypothesis). We should add that survival studies usually involve hazard ratios that are of course statistically very similar to odds ratios. Weighted average effects can be calculated from various hazard ratios in a meta-analysis of Cox regression studies.

#### Calculation of point estimates and their variances

Contingency table	numbers of patients with disease improvement	numbers of patients with no improvement	total
test treatment	a	b	a+b
reference treatment	c	d	c+d
total	a+c	b+d	n

Point estimators include relative risk (RR), odds ratio (OR), or risk difference (RD):

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

$$OR = \frac{a/b}{c/d}$$

$$RD = \frac{a}{(a+b)} - \frac{c}{(c+d)}$$

The data can be statistically tested by use of a chi-square test of the added point estimators.

Instead of RR and OR we take lnRR and lnOR in order to approximate normality

$$\text{Chi-square} = \frac{\left( \frac{\ln\text{RR}_1}{s_1^2} + \frac{\ln\text{RR}_2}{s_2^2} + \frac{\ln\text{RR}_3}{s_3^2} \dots \right)^2}{\frac{1}{s_1^2} + \frac{1}{s_2^2} + \frac{1}{s_3^2} + \dots}$$

degrees of freedom 1 (one).

$s^2$  = variance of point estimate :

$$s_{\ln\text{RR}}^2 = 1/a - 1/(a+b) + 1/c - 1/(c+d)$$

$$s_{\ln\text{OR}}^2 = 1/a + 1/b + 1/c + 1/d$$

$$s_{\text{RD}}^2 = ab/(a+b)^3 + cd/(c+d)^3$$

for RD, which does not have so much skewed a distribution, ln-transformation is not needed.

$$\text{Chi-square} = \frac{\left( \frac{\text{RD}_1}{s_1^2} + \frac{\text{RD}_2}{s_2^2} + \frac{\text{RD}_3}{s_3^2} \dots \right)^2}{\frac{1}{s_1^2} + \frac{1}{s_2^2} + \frac{1}{s_3^2} + \dots}$$

As alternative approach to the *weighted average effect analysis* the Mantel-Haenszel-(MH) summary chi-square can be used. It is a stratified analysis, which means that studies are analyzed as strata. The analysis' results is, though, pretty similar to those of the weighted average effect analysis. We should add, that the MH summary chi-square test is computationally much more complex, and that a computer is badly needed for the purpose:

$$\chi^2_{\text{M-H}} = \frac{\left( \sum a_i - \sum [(a_i + b_i)(a_i + c_i)/(a_i + b_i + c_i + d_i)] \right)^2}{\sum [(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)/(a_i + b_i + c_i + d_i)^3]}$$

$a_i$ ,  $b_i$ ,  $c_i$ , and  $d_i$  are the a-value, b-value, c-value, and d-value of the  $i$ th sample.

A second alternative is Peto's z-test, sometimes called Peto odds ratio.

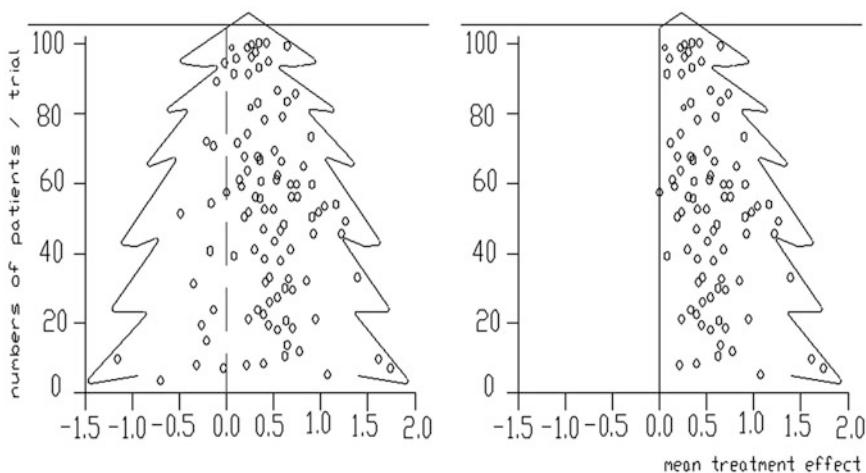
$$z = \frac{\sum (a_i - [(a_i + b_i)(a_i + c_i)/(a_i + b_i + c_i + d_i)]/n_i)}{\sqrt{[\sum (a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)/a_i + b_i + c_i + d_i]^2(a_i + b_i + c_i + d_i - 1)]}}$$

Results of the three approaches yield similar results. However, with Mantel-Haenszel and Peto the calculation of pooled variances is rather complex, and a computer program is required. Peto tends to cause over- and underestimation of extreme values like odds ratios of odds ratios  $>5$  or  $<0.2$ , but a good performance is generally obtained with rare events (Yusuf, Peto, et al, Beta blockade during and after myocardial infarction an overview of randomized trials, Progr Cardiovasc Trials 1985; 27: 355–71). A good starting point with any statistical analysis is plotting the data.

We should add that, in addition to means of continuous data, regression estimators like regression coefficients and correlation coefficients may serve as endpoint estimators in meta-analyses. Indeed, in treatment comparisons, e.g. new treatment versus control or placebo, a regression analysis may be used with treatment modality as predictor and treatment outcome as dependent variable. The magnitude of the regression coefficients of treatment modality versus outcome and that of the correlation coefficient are adequate measures to estimate the treatment efficacy of a new treatment. In meta-analysis a weighted mean of the  $b$  or  $r$  coefficients can be used for overall assessment.

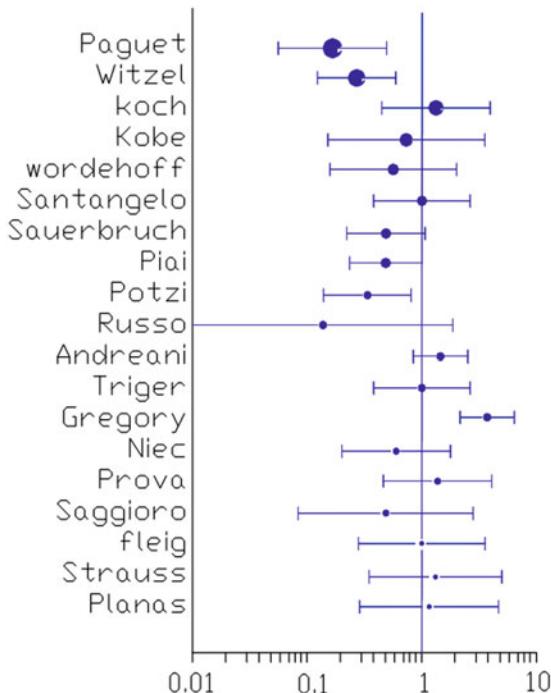
## 1.8 First Pitfall

We will now address the pitfalls. An important part of the data analysis is checking the pitfalls.



The first pitfall is the presence of significant publication bias in your meta-analysis. The above figure gives an example. On the x-axis are the results of the studies, on the y-axis the sample sizes of the studies. A statistical necessity would be that small studies have a large variance, large studies a small one. Also, a statistical necessity would be a pretty symmetrical pattern. A cut-off pattern like the one above indicates that a number of studies has not been published. Publication bias literally means, that small studies with a negative result have not been published. In order to avoid publication bias published and unpublished data should be included.

## 1.9 Second Pitfall

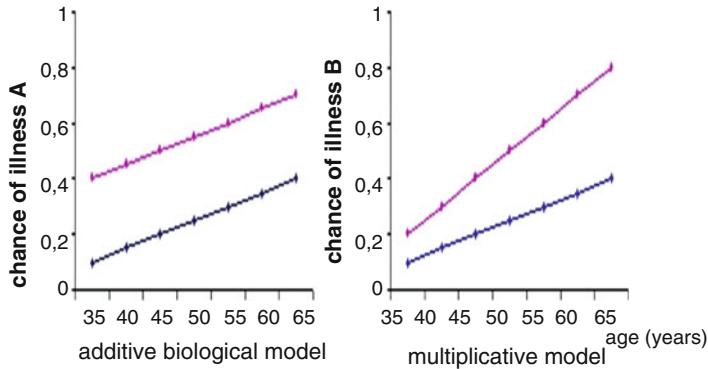


The second pitfall is heterogeneity. The above example shows on the x-axis the odds ratios of 19 studies. Odds ratio (OR) means here the chance of fatal oesophageal varices bleeding with sclerotherapy/without sclerotherapy. If this ratio is  $<1$ , then sclerotherapy will be helpful, if  $>1$ , then it will be not so. Testing heterogeneity means testing, whether the differences between the main outcomes of the studies are larger than could occur by chance (multiple groups analysis of variance (ANOVA), or Chi-square tests are used for the testing).

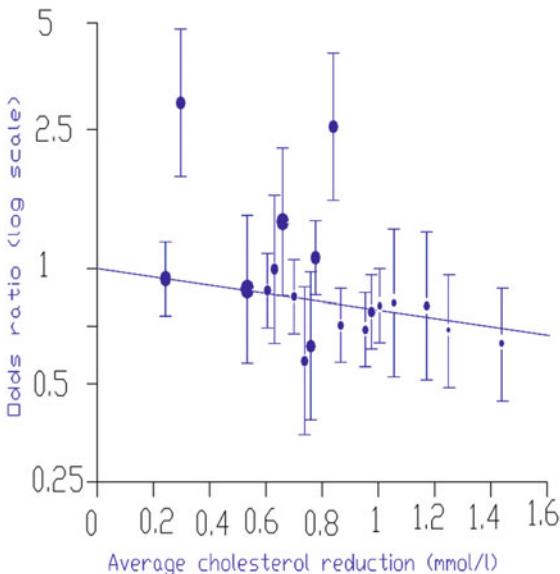
How do you test heterogeneity? For continuous data multiple groups ANOVA is adequate. Assess whether the between-study variance is larger than within-study variance. For odds ratios (ORs), risk ratios (RRs), or risk differences (RDs) of binary data multiple groups chi-square tests can be used. Assess whether the variance between the ORs is large compared to the variance in their 95% confidence intervals (CIs). Normalize the ORs, because CIs around ORs are skewed. An example of the fixed effect calculation of RDs is given underneath ( $s^2$  = variance, the  $\chi^2$  test has  $n-1$  degrees of freedom).

$$\chi^2 = \frac{RD_1^2}{s_1^2} + \frac{RD_2^2}{s_2^2} + \frac{RD_3^2}{s_3^2} \dots - \frac{\left[ \frac{RD_1}{s_1^2} + \frac{RD_2}{s_2^2} + \frac{RD_3}{s_3^2} \dots \right]^2}{\frac{1}{s_1^2} + \frac{1}{s_2^2} + \frac{1}{s_3^2} + \dots}$$

A random effect model for heterogeneity of Dersimonian and Laird assumes heterogeneity due to some unexpected subgroup effect, and introduces a separate random variable for the purpose. If both traditional and random effect analyses are not statistically significant, then no heterogeneity is in the meta-analysis.



If, however, significant heterogeneity is in a meta-analysis, this will not be a disaster. You must, though, find the cause(s). For example in the above figure it is shown that the chance of disease will get smaller if age gets lower. The right hand graph shows heterogeneity because of a nonlinear relationship here.



Another cause of heterogeneity is shown in the above figure, where the chance of myocardial infarction will be reduced if cholesterol is reduced. This relationship is disturbed by outliers. The differences in outcome between the studies is due to two severe outliers.

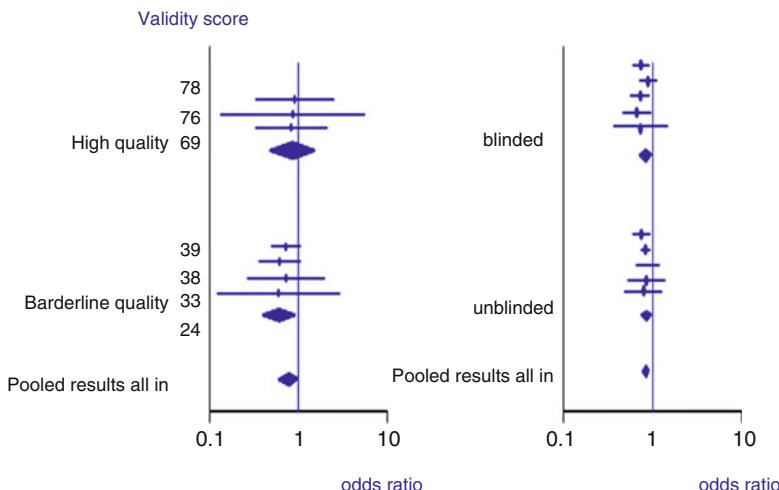
	No of cancer cases	No of studies	Mean (SE) difference in serum cholesterol (mmol/l)*	Heterogeneity
<hr/>				
All studies:				
Overall	12 516	33	-0.041 (0.009)	$\chi^2=53$ , df=32, P=0.01
Socioeconomic status:				
High	619	4	+0.032 (0.048)	
Mixed	10 378	20	-0.030 (0.010)	$\chi^2=37$ , df=30, P=0.18
Low	1 519	9	-0.130 (0.025)	

\* Mean cholesterol in those who subsequently developed cancers minus mean in those who did not.

A third cause of heterogeneity is given in the above table. It shows outlier-populations. The chance of cancer will be reduced, if cholesterol is reduced.

However, only in the lowest-social-group this is significantly-so, and, the heterogeneity is, thus, probably due to a confounding social factor.

## 1.10 Third Pitfall



A lack of robustness is a third pitfall. Robustness is synonymous to sensitivity. Sensitivity analysis is the study of how the uncertainty in the output of a statistical analysis can be assigned to different sources of uncertainty in the input. In meta-analysis it is often defined as the phenomenon, that studies with borderline quality produce more spectacular results than high quality studies do. It is usually caused by placebo-effects, also doctor-mediated placebo-effects. What to do? As shown in the above figure, we remove the low quality studies, and, subsequently, look, whether pooling will show changes in the results. If so, don't pool. Leaving out studies at this stage is scientifically impossible (after inclusion, exclusion is no more possible).

## 1.11 Benefits and Criticisms of Meta-analyses

A major purpose of all scientific research is establishing causal rather than confounding relationships. Meta-analyses are a current strategy for assessing a large volume and diversity of data for the purpose. They aim at estimating a common pooled effect with increased precision. They also aim at evaluating diversity by exploring and explaining heterogeneities between studies. They should

perform better than the already pretty good 60s *causality rules* from London UK statistician Bradford Hill (Proc Roy Soc Med 1965; 58: 295–300).

1. Strength of association between variables.
2. Consistency between studies.
3. High specificity of predictor variables.
4. Temporality.
5. Dose-response patterns.
6. A plausible biological mechanism.
7. Coherence of lab and in vivo findings.
8. Controlled experiments.
9. Analogous cause effect relationships have been found before.

Nonetheless, the rules are subjective, and biases can not be excluded. Another current methodology for finding causal relationships includes machine learning methodologies, particularly Bayesian networks (Machine learning in medicine part two, Chap. 16 Bayesian networks, 163–70, Springer Heidelberg Germany, 2013). Relationships are often by chance/accident. But, if you find mathematically a relationship between 3 instead of 2 factors, then the chance of a causal mechanism will be a bit more probable. Bayesian network uses a trick: standardized regression coefficients between 3 or more factors can be, simply, added up. It is, e.g., the basis of structural equation modeling in the SPSS AMOS software package. Although very successful by biomolecular scientists for describing novel pathogenic and metabolic pathways, it is somewhat more problematic with biomedical data. Causality is, notoriously, difficult with observational data, and biases due to selection and other multivariate factors require cross validations for reliability assessments.

This chapter is the first of an edition focusing on meta-analyses as a method for establishing scientific standards of evidence. A core advantage is, that they are governed by the traditional rules for scientific research, including those of scientific rigor (current chapter), and of the scientific method (Chap. 3). Yet, meta-analyses have been criticized for not adequately predicting the results of subsequent large trials. Also, they are not good at predicting serious (and non-serious) adverse effects. Why would that be so? Probably, it is due to the above three pitfalls.

Initiatives against pitfalls like the statement as given by the 70 journal editors of the CONSORT group (Consolidated Standards of Reporting Trials (CONSORT), and others like the Unpublished Paper Amnesty Movement, and the World Association of Medical Editors (see Understanding Clinical Data Analysis, Chap. 3, 2016, Springer Heidelberg Germany, from the same authors) will be helpful. The Cochrane-Group/Evidence-based Movement has offices in every western country, and is very much in favor of meta-analyses as the golden standard in clinical research. This golden standard includes routine functions in the field of scientific research like the following.

1. Reporting randomized experimental research requirement.
2. Development of new drugs.
3. Determination of individual therapies.

4. Leading the way for regulatory organs.
5. Epidemiological research.



New developments include meta-analyses purely explorative in nature. The usual primary question “is result representative” is here replaced with plenty of information regarding subgroups and interactions (see also Primer in statistics meta-analysis, Circulation 2007; 115: 2870–5, from the same authors). These kind of explorative meta-analyses are like technological evaluation reports of physicists.

The trend of explorative meta-analyses is enhanced by the internet, which of course registers many more data than the traditional medical journals did. In 2004 the J Am Med Assoc and the Lancet published 9 meta-analyses each, all of them hypothesis-driven, in the same year the high impact journal for cardiovascular state of the art Circulation published 16 meta-analyses, 8 of them purely explorative in nature.

## 1.12 Conclusion

The current chapter was written for physicians and other health workers, non-mathematicians, and for their benefit the authors refrained from using mathematical equations.

### Meta-analyses

- are secondary, otherwise called post/hoc, analyses,
- test, however, primary, not secondary, hypotheses,

- have probability statements, that may, therefore, be as valid as those of the primary studies.

A meta-analysis is a systematic review of trials with pooled data. How long do they exist? In the early 70s psychologists were the first to perform systematic reviews, without pooled data. Only pretty recently, since 1995, more homogeneous trials were published, and more pooling could take place. Why do we perform meta-analyses? This is pretty obvious: with more data, we are likely to obtain certainty. However, also more differences between subgroups may be observed.

The essentials of meta-analyses need just a few words. All that is needed, is the rules of scientific rigor and a brief list of pitfalls. Scientific rigor requires that we stick to

- (1) a clearly defined prior hypothesis,
- (2) to a thorough search of trials,
- (3) to strict inclusion criteria for trials, and
- (4) to uniform guidelines for data analysis.

The brief list of pitfalls includes

- (1) checking publication bias,
- (2) checking heterogeneity,
- (3) checking robustness.

Why use effect sizes instead of significance tests in meta-analyses (Neil, Meta-analysis Research Methodology, May 10 2006, [www.sciencepublishinggroup.com](http://www.sciencepublishinggroup.com)). 2006). You cannot summarize quantitatively the results of different quantitative studies with the help of p-values only. Instead, weighted means and their weighted standard errors are used to quantify overall effects sizes and pooled heterogeneity. Ultimately, these results can be tested using fairly simple tests like t-tests and chi-square tests.

## Reference

More background, theoretical and mathematical information of meta-analyses is given in Statistics applied to clinical studies 5th edition, Chaps. 32–34 and 48, Springer Heidelberg Germany, 2012, from the same authors.

# **Chapter 2**

## **Mathematical Framework**

### **Working Papers with Emphasis on Entire Data Coverage**

**Abstract** In a meta-analysis weighted averages are computed for the purpose of establishing, whether scientific findings are consistent, and can be generalized across populations and treatment variations, and whether findings vary between subgroups.

Continuous data are summarized with means and standard deviations. Binary data are averaged with risk differences, relative risks, odds ratios. Survival data are summarized with Kaplan Meier curves and hazard ratios. Publication bias is assessed with funnel plots, otherwise called Christmas trees, and the shift of risk ratios caused by the addition of unpublished trials from abstract-reports and proceedings of scientific meetings. Heterogeneity is assessed with fixed and random effect tests, and with the I-square- statistic. Sensitivity is assessed with high and low quality studies analyzed separately. Novel developments are reviewed.

### **2.1 Introduction**

In the previous chapter a nonmathematical review of meta-analysis methodologies has been given. In this chapter we will give the mathematical framework of it. Meta-analyses can be defined as systematic reviews with pooled data. Traditionally, they are post-hoc analyses. However, probability statements may be more valid, than they usually are with post-hoc studies, particularly if performed on outcomes that were primary outcomes in the original trials. Problems with pooling are frequent: correlations are often nonlinear; effects are often multifactorial rather than unifactorial; continuous data frequently have to be transformed into binary data for the purpose of comparability; poor studies may be included and coverage may be limited; data may not be homogeneous and may fail to relate to hypotheses. In spite of these problems, the methods of meta-analysis are an invaluable scientific activity: they establish whether scientific findings are consistent, and can be generalized across populations and treatment variations, and whether findings vary between subgroups. The methods also limit bias, improve reliability and accuracy of conclusions, and increase the power and precision of treatment effects and risk exposures. We first will review the statistical analysis, including the analysis of potential pitfalls. Finally, we will cover some new developments.

## 2.2 General Framework

In general, meta-analysis refers to statistical analysis of the results of different studies. The simplest analysis is to calculate an average, and in a meta-analysis a weighted average is computed. Consider a meta-analysis of  $k$  different clinical trials, and let  $x_1, x_2, \dots, x_k$  be the summary statistics. The *weighted average effect* is then calculated as

$$\bar{X}_w = \frac{\sum_{i=1}^k w_i x_i}{\sum_{i=1}^k w_i}$$

and its standard error is

$$se(\bar{X}_w) = \left[ \frac{\sum_{i=1}^k (w_i)^2 Var(x_i)}{\left[ \sum_{i=1}^k w_i \right]^2} \right]^{1/2}.$$

The weights  $w_i$  are a function of the standard error of  $x_i$ , here denoted as  $se(x_i)$ , and of the variance  $\sigma^2$  of the true effects of the new medicine between  $k$  different studies:

$$w_i = \frac{1}{(se(x_i))^2 + \sigma^2}.$$

- If all of the  $k$  studies have the same true quantitative effect, then the variance  $\sigma^2$  will equal 0, and the weighted average effect will be called a *fixed-effect* estimate.
- If the true effects of the compound vary between studies, then the variance  $\sigma^2$  will be larger than 0, and the weighted average effect will be called a *random-effect* estimate.

For the fixed-effect estimate (meaning that the variance  $\sigma^2 = 0$ ) the calculations are quite simple, for the random-effect estimate the calculations are more complex, but available in computer packages (SAS, SPSS Statistical Software, Cochrane Revman, Stata statistical software for professionals, Comprehensive Meta-analysis from Biostat 2011, etc.). In the Chaps. 4 and 6 some random-effect tests on a pocket calculator will be given for the benefit of those who are not mathematician, and, yet, wish to completely understand what they are doing.

Depending on the type of variables, the main outcome values of the different studies, often called summary statistics, and here called  $x_1, x_2, \dots, x_k$ , have different forms.

## 2.3 Continuous Outcome Data, Mean and Standard Deviation

Continuous data are summarized with means and standard deviations.

### 2.3.1 Means and Standard Deviation (SD)

$\text{mean}_{1i}$  and  $\text{SD}_{1i}$  in the placebo-group

$\text{mean}_{2i}$  and  $\text{SD}_{2i}$  in the active-treatment group of trial i.

The summary statistic equals  $x_i = \text{mean}_{1i} - \text{mean}_{2i}$  and its standard error

$$\text{se}(x_i) = \sqrt{\frac{\text{SD}_{1i}^2}{n_{1i}} + \frac{\text{SD}_{2i}^2}{n_{2i}}},$$

where  $n_{1i}$  and  $n_{2i}$  are the sample sizes of the two treatments.

**Note 1** if a trial compares two treatments in the same patients, we have a crossover trial, and the summary statistic is  $x_i = \text{mean}_{1i} - \text{mean}_{2i}$ , where

$\text{mean}_{1i}$  and  $\text{mean}_{2i}$  = means of the two treatments, and its standard error

$$\text{se}(x_i) = \sqrt{\frac{\text{SD}_{1i}^2}{n_i} + \frac{\text{SD}_{2i}^2}{n_i} - \frac{2 r \text{SD}_{1i} \text{SD}_{2i}}{n_i}},$$

where  $r$  is the correlation between the outcomes in the two treatments.

**Note 2** if the distribution of the outcomes is very skewed, it is often more useful to summarize outcomes with medians than means.

## 2.4 Continuous Outcome Data, Strictly Standardized Mean Difference (SSMD)

Current meta-Analyses are often analyzed with standardized mean differences, otherwise called “strictly standardized mean difference” (SSMD) as effect size, instead of means and mean differences. SSMD is sometimes also called Cohen’s d. It is not equal to a t-value.

$$\text{Standardized mean difference} = \frac{\text{mean}_{\text{treatment 1}} - \text{mean}_{\text{treatment 2}}}{\text{SD}_{\text{treatment 1}}^2 + \text{SD}_{\text{treatment 2}}^2}$$

$$t - \text{value} = \frac{\text{mean}_{\text{treatment 1}} - \text{mean}_{\text{treatment 2}}}{\text{se}_{\text{treatment 1}}^2 + \text{se}_{\text{treatment 2}}^2}$$

Remember: samples of continuous data are characterized by

$$\text{Mean} = \frac{\sum x}{n} = \bar{x}$$

where  $\Sigma$  is the summation, x are the individual data, n is the total number of data.

$$\text{Variance} = \sum (x - \bar{x})^2$$

$$\text{Mean variance} = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Mean variance is often briefly named variance. And, so, don’t forget the term variance is commonly used to name mean variance. The famous term standard deviation is often abbreviated as, simply, s, and is equal to the square root of this mean variance.

$$\text{standard deviation (SD)} = \sqrt{(\text{mean variance})}$$

For statistical testing SDs are not convenient, but standard errors (se-values) are very much so, and

$$\text{se} = \text{SD} / \sqrt{n}.$$

The weighted average effect as mentioned above using SSMDs as effect sizes of the individual studies in a meta-analysis is given (\* = symbol of multiplication):

$$\text{weighted average effect} = \frac{\sum w_{\text{individual studies}} * \text{SSMD}_{\text{individual studies}}}{\sum w_{\text{individual studies}}}$$

$$\text{se}_{\text{weighted average effect}} = \frac{\sqrt{\sum w_{\text{individual studies}}^2} * \text{Variance}_{\text{individual studies}}}{(\sum w_{\text{individual studies}})^2}$$

$$w_{\text{individual studies}} = 1 / (\text{se}_{\text{individual studies}})^2$$

With SSMDs as effect sizes of the studies included in a meta-analysis, and n of the patients receiving treatment 1 and treatment 2 identical, then they can be calculated:

$$\text{SSMD}_{\text{individual study}} = \frac{\text{mean}_{\text{treatment 1}} - \text{mean}_{\text{treatment 2}}}{\sqrt{(\text{SD}^2_{\text{treatment 1}} + \text{SD}^2_{\text{treatment 2}})}}$$

The above enumerator term  $\sqrt{\sum w^2_{\text{individual studies}}} * \text{Variance}_{\text{individual studies}}$  is very interesting, because

$$w^2_{\text{individual studies}} * \text{Variance}_{\text{individual studies}}$$

reduces per study to

$$[1/(se^2_{\text{treatment 1}} + se^2_{\text{treatment 2}})].[(\text{SD}^2_{\text{treatment 1}} + \text{SD}^2_{\text{treatment 2}})].$$

with the sample sizes “n” of treatment 1 and treatment 2 equal, and  $se = SD/\sqrt{n}$ , it eventually will even reduce to:

$$n * 1 \text{ with } * \text{ again as symbol of multiplication.}$$

Obviously, the larger the study, the more its effect size will contribute to the meta-analyzed effect size.

## 2.5 Continuous Outcome Data, Regression Coefficient and Standard Error

Instead of mean and standard error, a regression coefficient, otherwise called b-value, and its standard error can be applied. The mean result of a parallel-group study can be readily expressed in the form of a b-value and its standard error. This involves linear regression analysis. In a linear regression equation  $y = a + bx$ ,  $b = \text{the covariance of the } x \text{ and } y \text{ values divided by the squared standard deviation of the } x\text{-values}$ , but can in the event of binary predictors, namely treatment 0 and treatment 1 easily be computed with the help of the terms  $\text{Mean}_{1i}$  and  $\text{SD}_{1i}$  in the placebo-group, and  $\text{mean}_{2i}$  and  $\text{SD}_{2i}$  in the active-treatment group of trial i .

The regression coefficient equals the summary statistic  $x_i = \text{mean}_{1i} - \text{mean}_{2i}$ , and its

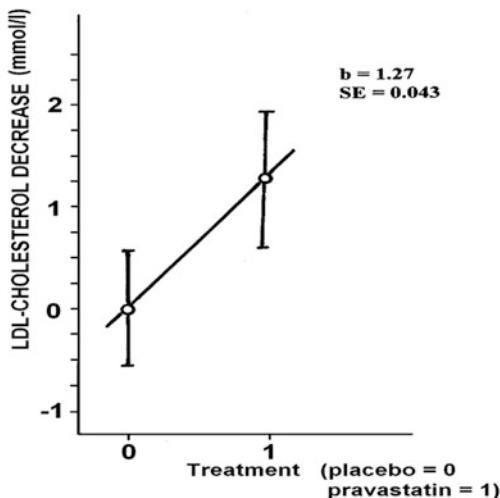
$$\text{pooled standard error equals is } \sqrt{\frac{\text{SD}_{1i}^2}{n_{1i}} + \frac{\text{SD}_{2i}^2}{n_{2i}}}$$

As an example a parallel-group study with pravastatin versus placebo for the treatment of cholesterol from one of our group (REGRESS study, Circulation 1995; 91: 2528–40):

placebo	pravastatin	difference	
n	434	438	
mean	-0.04	1.23	1.27
SD	0.59	0.68	SEM = 0.043

The same result is obtained by drawing the best-fit-regression-line for data.

The regression coefficient is equal to the difference of the means, its standard error is equal to the standard error of the difference of the means, namely 1.27 and 0.043 mmol/l.



## 2.6 Continuous Outcome Data, Student's T-Value

In controlled trials usually the differential effect of two treatments is assessed. The t-test is adequate for testing. Meta-analyses of multiple similar studies is performed with pooling.

We just take the mean result of the mean difference of the outcome variable we want to meta-analyze and add up. The data can be statistically tested according to unpaired t-test of the sum of multiple means:

$$t = \frac{\text{mean}_1 + \text{mean}_2 + \text{mean}_3}{\sqrt{\text{SEM}_1^2 + \text{SEM}_2^2 + \text{SEM}_3^2 + \dots}} \text{ with degrees of freedom} \\ = n_1 + n_2 + n_3 + n_k - k$$

$n_i$  = sample size ith sample,  $k$  = number of samples, SEM = standard error of the mean.

If the standard deviations are very different in size, e.g., if one is twice the other, then a more adequate calculation of the pooled standard error is as follows. This formula gives greater weight to the pooled SEM the greater the samples.

$$\text{Pooled SEM} = \sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2 + \dots}{n_1 + n_2 + \dots - k} \times \left( \frac{1}{n_1} + \frac{1}{n_2} + \dots \right)}$$

Similarly, if the samples are very different in size, then a more adequate calculation of the nominator of t is as follows.

$$k \left( \frac{\text{mean}_1 n_1 + \text{mean}_2 n_2 + \dots}{n_1 + n_2 + \dots} \right)$$

## 2.7 Continuous Outcome Data, Correlation Coefficient (R or r) and Its Standard Error

R = the correlation coefficients, and is equal to the covariance of x and y values divided by the square root of the product of the variance of the x values and that of the y values. In studies reporting correlation coefficients without measure of spread, the standard error can be estimated with the equation

$$\text{standard error} = \sqrt{[(1 - r^2)/(n - 2)]}.$$

As this estimate has been recognized to underestimate the true standard error of the data, it may be wise to replace it with another approximation procedure, *the r to z transformation of Fisher*. For that purpose it is assumed that not only y-data follow identical normal frequency distributions, but also x-data do so. The correlation coefficients are replaced with z-values obtained from the underneath equation is given by ( $\log$  = natural log = naperian log =  ${}^e\log$ ). The maths make use of the delta method and are pretty complex and beyond the current text.

$$z = \log \sqrt{[(1 + r)/(1 - r)]}$$

The standard error of this z-value is approximately equal to the square root of  $n$ ,  $n-3$  produces a slightly better precision of estimates:

$$se = 1/\sqrt{n-3}.$$

$n$  = Sample size of study. The z-transformation can be used for the computation of the weighted r and its standard error of a meta-analyses of r-values. Let us assume

4 studies have produced the r-values  $r_1$  to  $r_4$  and they have variances of  $se^2_1$  to  $se^2_4$ . Then the following computations can be made.

$\log \sqrt{[(1+r)/(1-r)]} = z$  transformation version of the averaged r from the studies in the meta-analysis.

$$se^2_{r \text{ average}} = \frac{(1 - r^2_1)^2}{n_1 - 3} + \frac{(1 - r^2_2)^2}{n_2 - 3} + \frac{(1 - r^2_3)^2}{n_3 - 3} + \frac{(1 - r^2_4)^2}{n_4 - 3},$$

where  $r_1$  = correlation coefficient of first study, and  $n_1$  = sample size of first study etc., and  $r_{\text{average}}$  = is the unweighted average of the separate r-values from the different studies, and the z-transformation  $r_{\text{average}} = \log \sqrt{[(1 + r_{\text{average}})/(1 - r_{\text{average}})]}$ .

$$se^2 \text{ of the } z \text{-transformation of } r_{\text{average}} = \frac{se^2_{r \text{ average}}}{(1 - r^2_{\text{average}})^2}$$

$$\begin{aligned} 95\% \text{ confidence of } z\text{-transformation of } r_{\text{average}} &= z\text{-transformation of } r_{\text{average}} \\ &\pm 2 \left[ \frac{(se^2_{r \text{ average}})^{1/2}}{(1 - r^2_{\text{average}})^2} \right] \end{aligned}$$

$z\text{-transformation of } r_{\text{average}}$  divided by  $\frac{se^2_{r \text{ average}}}{(1 - r^2_{\text{average}})^2}$  produces a z-value.

If  $z = 2$ , then  $p = 0.05$ , meaning that the a-transformed  $r_{\text{average}}$  will be statistically significant at  $p = 0.05$ , and, thus, that it, thus, will be significantly better than a z-transformed  $r_{\text{average}}$  of zero.

Likewise, if the 95% confidence interval of the z-transformed  $r_{\text{average}}$  is between e.g., 0.6 and 0.9, then we can be 95% certain that next time a similar meta-analysis will provide a z-transformed  $r_{\text{average}}$  between 0.6 and 0.9.

A hypothesized data example will be given. In patients with chronic constipation a novel laxative was tested against the standard laxative bisacodyl with numbers of stool per month as main outcome. Four studies at different sites were performed and the novel laxative was generally better than the standard. However, the investigators were particularly interested to find out, whether or not a positive correlation would exist between the effect of the new laxative and the control treatment, in other words whether the poorest responders to the standard treatment would also be the poorest responders to the new treatment.

The four linear correlation coefficients (r values) of new versus standard laxative were

$$0.79 \quad 0.50 \quad 0.70 \quad 0.56$$

samples size of studies

$$35 \quad 160 \quad 65 \quad 30$$

$$r_{\text{average}} = (0.79 + 0.50 + 0.70 + 0.56)/4 = 0.64$$

$$\begin{aligned} z - \text{transformation of } r_{\text{average}} &= \log\sqrt{[(1+r_{\text{average}})/(1-r_{\text{average}})]} \\ &= \log\sqrt{(1.64/0.36)} = 0.76 \end{aligned}$$

$$se^2_{r_{\text{average}}} = \frac{(1-0.79^2)^2}{32} + \frac{(1-0.50^2)^2}{157} + \frac{(1-0.70^2)^2}{62} + \frac{(1-0.56^2)^2}{27} = 0.0299$$

$$se^2 \text{ of the } z - \text{transformed } r_{\text{average}} = 0.0299/(1-0.64^2)^2 = 0.0859$$

$$\begin{aligned} 95\% \text{ confidence interval of the } z\text{-transformed } r_{\text{average}} &= 0.76 \pm 2*0.0852^{1/2} \\ &= 0.76 \pm 2*0.29 \\ &= \text{between 0.47 and 1.00} \end{aligned}$$

The meta-analyzed correlation coefficient of 0.76 with  $se = 0.29$  is significantly larger than a correlation coefficient of 0.0 at  $p = 0.010$ , because  $z = 0.76/0.29 = 2.621$ .

The traditional t-test of the average  $r$  with  $se = \sqrt{[(1-r^2)/n-2]}$  would produce a better p-value,  $p = 0.0001$ , because  $0.64/\sqrt{[(1-r^2)/n-2]} = 0.64/0.038 = 16.84$ . But, it does not appropriately account the weights of the  $r$  values of the separate studies in the meta-analysis, and underestimates the standard error of the average correlation. The  $r$  to  $z$  transformed approach to testing meta-analyses of studies with correlation coefficients as outcome may produce slightly less sensitivity of testing, but, instead, heterogeneity of correlation coefficients of the different studies is better taken into account.

## 2.8 Continuous Outcome Data, Coefficient of Determination $R^2$ or $r^2$ and Its Standard Error

$R$ -square =  $R^2$  is not only the square of the standardized regression coefficient, but also the proportion of certainty provided by the  $x$ -values about the  $y$ -values. E.g., with a standardized regression coefficient of 0.8, the  $r$ -square will equal 0.64. We will be 64% certain about the value of  $y$  if we know the value of  $x$ . A standard error needs to be added if we wish to test whether the percentage certainty is significantly better than zero percent. Otherwise,  $z$ -transformed  $r_{\text{average}}$  can be applied likewise in a meta-analysis reported with  $r$ -values as main outcome. E.g., in the above example, we can predict the efficacy of the new laxative with  $0.76^2 = 0.58 = 58\%$  certainty, if we know the efficacy of the control laxative. The use of  $r$  square has some disadvantages.

First, it does not tell us whether the correlation between the x and y values are positive or negative correlated. It is a goodness of fit test for the linear model, and an r-square of 1 (= 100%) means a perfect fit.

Second, you can be sure about the magnitude of your y-value, if you know the size of the associate x-value. However, with multiple x-variables, and a single y-variable we have a multiple linear regression, and with increasing numbers of x-variables the r-square increasingly rises, because multiple x-variables tend to give more information about the y-variable than a single one. This causes r-squares to rise, and to provide better certainty about the outcome, the y-values. This phenomenon is pejoratively called kitchen sink regression, and is an important bias of regression analyses producing excellent results due to the play of gambling rather than a real probabilistic effect (it is like data dredging, see, e.g., Chap. 1 Statistics applied to clinical studies, Springer Heidelberg Germany, 2012, from the same authors).

Nonetheless, r-squares are beautiful, particularly with meta-analyses. The multiple studies are not tested more than a single time, and the r-squares better discriminate than usual r-values, because they have a squared linear rather than simple linear relationship.

## 2.9 Binary Outcome Data, Risk Difference

Binary data are summarized as proportions of patients with a positive outcome in the treatment arms, which can be denoted as  $p_{1i}$  and  $p_{2i}$  for the treatments 1 and 2. Three different summary statistics are routinely used. The summary statistic of trial i equals  $x_i = p_{1i} - p_{2i}$ , the standard error equals

$$se(x_i) = \sqrt{\frac{p_{1i}(1 - p_{1i})}{n_{1i}} + \frac{p_{2i}(1 - p_{2i})}{n_{2i}}},$$

where  $n_{1i}$  and  $n_{2i}$  are the sample sizes of the two treatments of trial i.

## 2.10 Binary Outcome Data, Relative Risk

The summary statistic of trial i equals the ratio of the two proportions, but its distribution is often very skewed. Therefore, we prefer to analyze the natural logarithm of the relative risk,  $\ln(RR)$ . The summary statistic thus equals

$$x_i = \ln(p_{1i}/p_{2i}),$$

and the standard error equals

$$\text{se}(x_i) = \sqrt{\frac{1 - p_{1i}}{p_{1i}n_{1i}} + \frac{1 - p_{2i}}{p_{2i}n_{2i}}}$$

## 2.11 Binary Outcome Data, Odds Ratio

The summary statistic of trial  $i$  equals the ratio of the odds, but since the odds ratio is strictly positive, we again prefer to analyze the natural logarithm of the odds ratio. Thus the summary statistic equals

$$x_i = \ln \left( \frac{p_{1i}/(1 - p_{1i})}{p_{2i}/(1 - p_{2i})} \right).$$

and the standard error equals

$$\text{se}(x_i) = \sqrt{\frac{1}{n_{1i}p_{1i}} + \frac{1}{n_{1i}(1 - p_{1i})} + \frac{1}{n_{2i}p_{2i}} + \frac{1}{n_{2i}(1 - p_{2i})}}.$$

**Note** the Mantel-Haenszel method has been developed for a stratified analysis of odds ratios, and has been extended to the stratified analysis of risk ratios and risk differences. Like the general model, a weighted average effect is calculated. For the calculation of combined odds ratios Peto's method is also often used. It applies a way to calculate odds ratios, which may cause under- or overestimation of extreme values like odds ratios  $<0.2$  or  $>5.0$ .

**Note** sometimes valuable information can be obtained from crossover studies, and, if the paired nature of the data are taken into account, such data can be included in a meta-analysis. The Cochrane Library CD-ROM provides the Generic inverse variance method for that purpose.

## 2.12 Binary Outcome Data, Survival Data

Survival trials are summarized with Kaplan-Meier curves, and the difference between the survival in two treatment arms is quantified with the log (hazard ratio), as calculated from the Cox regression model. To test whether the weighted average is significantly different from 0.0, a chi-square test is used.

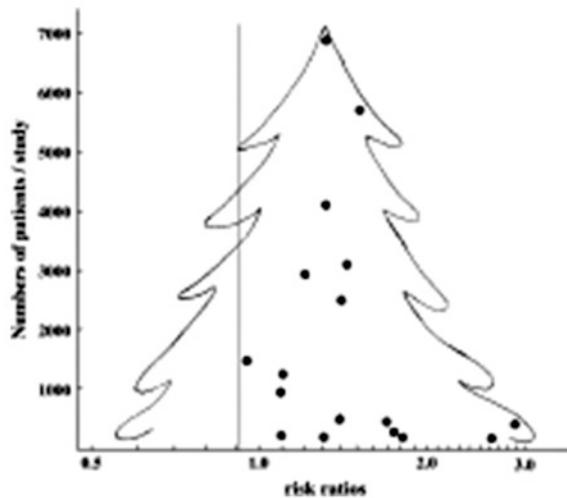
$$\chi^2 = \left( \frac{\bar{X}_w}{se(\bar{X}_w)} \right)^2 \text{ with one degree of freedom.}$$

**Note** a calculated  $\chi^2$ -value larger than 3.841, indicates that the pooled average is significantly different from 0.0 at  $p < 0.05$ , and, thus, that a significant difference exists between the test and reference treatments. The Generic inverse variance method is also possible for the analysis of hazard ratios (hazard ratios are odds ratios of hazard data).

## 2.13 Pitfalls, Publication Bias

Meta-analyses will suffer from any bias, that the individual studies, as included, suffered from, including incorrect and incomplete data. As an example (1) out of 49 recently published studies, 83% of the nonrandomized and 25% of the randomized studies were partly refuted soon after publication (Ioannides, J Am Med Assoc 2005; 294: 210–28), (2) out of 519 recently published trials 20% selectively reported positive results, and reported negative results incompletely (Chan and Altman, Br Med J, 2005; 330: 753–6). Three common pitfalls of meta-analyses are listed.

A good starting point with any statistical analysis is plotting the data.

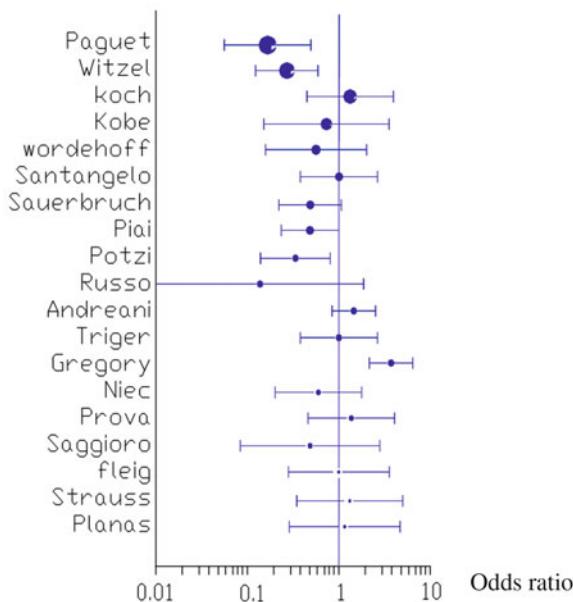


A Christmas tree or upside-down-funnel-pattern of distribution of the results of the published trials shows on the x-axis the risk ratio of each trial, on the y-axis the

sample size of the trials. The smaller the trial, the wider the distribution of results. Small studies with negative results are not published, and, thus, missing. This cut Christmas-tree can help suspect, that there is **publication bias** in the meta-analysis. Publication bias can be tested by calculating the shift of risk ratios caused by the addition of unpublished trials from abstract-reports or proceedings.

## 2.14 Pitfalls, Heterogeneity

In order to visually assess heterogeneity between studies several types of plots are proposed, including forest plots, radial and L' Abbe plots (National Council of Social Studies. Statistical and power analysis software. <http://www.ncss.com/metaanal.html>).



The above forest plot gives an example of a meta-analysis with odds ratios and 95% confidence intervals (CIs), telling something about heterogeneity. On the x-axis are the results, on the y-axis the sample size of the trials. We see the results of 19 trials of endoscopic intervention versus no intervention for upper intestinal bleeding: odds ratios less than one represent a beneficial effect. These trials were considerably different in patient-selection, baseline-severity-of-condition, endoscopic-techniques, management-of-bleeding-otherwise, and duration-of-follow-up. And so, this is a meta-analysis which is, clinically, very heterogeneous.

Is it also statistically heterogeneous? For that purpose we may use a fixed-effect model, which tests, whether there is a greater variation between the results of the trials, than is compatible with the play of chance, using a chi-square test. The null-hypothesis is that all studies have the same true odds ratio, and that the observed odds ratios vary only due to sampling variation in each study. The alternative hypothesis is, that the variation of the observed odds ratio is also due to systematic differences in true odds ratios between studies. The Cochran Q test with the Q statistic is used to test the above null hypothesis with summary statistics  $x_i$  and weights  $w_i$ :

$$Q = \sum_{i=1}^k w_i (x_i - \bar{X}_w)^2$$

with  $k-1$  degrees of freedom.

We find  $Q = 43$  for  $19-1 = 18$  degrees of freedom (dfs) for the example of the endoscopic intervention. The p-value is  $<0.001$  giving substantial evidence for statistical heterogeneity. For the interpretation it is useful to know, that, when the null-hypothesis is true, a Q statistic has on average a value close to the degrees of freedom, and increases with increasing degrees of freedom. So, a result of  $Q = 18$  with 18 dfs would give no evidence for heterogeneity, values much larger would do so for the opposite.

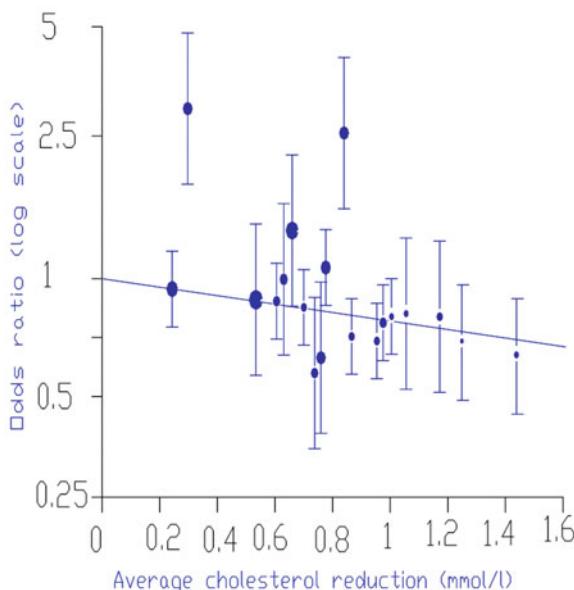
If the above test is positive, it is common to also calculate a random-effect estimate of the weighted average, as suggested by Dersimionian and Laird. We should add that, in most situations, the use of the random-effect model will lead to wider confidence intervals and a lower chance, that a difference is statistically significant. A disadvantage of the random-effect analysis is, that small and large studies are given almost similar weights. Additional information on random effect testing is given in the Chaps. 4 and 6. Complementary to the above Q-statistic, the amount of heterogeneity between studies is currently often quantified with the so-called  $I^2$ -statistic<sup>32</sup>

$$I^2 = 100\% * [Q - (k - 1)]/Q$$

which is interpreted, as the proportion of total variation in study estimates due to heterogeneity, rather than sampling error. Fifty % is, generally, used as a cut-off for heterogeneity.

When there is heterogeneity, careful investigation of the potential cause has to be accomplished. The main focus should be trying to understand any sources of heterogeneity in the data. Examples are given in the Chap. 1. In practice, it may be less hard to assess this, since the do-ers already have noticed clinical differences, and it, thus, becomes easy to test the data accordingly. The general approach is to quantify the association between the outcomes and characteristics of the different trials. Not only patient-characteristics, but also trial-quality-characteristics, such as the use of blinding, randomization, and placebo-controls have to be considered. Scatterplots are helpful to investigating the association between outcome and a

covariate, but these scatterplots must be inspected carefully, because differences in trial sample-sizes may distort the existence of association, and meta-regression techniques may be needed to investigate associations.



Outliers may also give a clue about the cause of heterogeneity. The above figure shows the relation between cholesterol and coronary heart disease. The two outliers on top were the main cause for heterogeneity in the data.

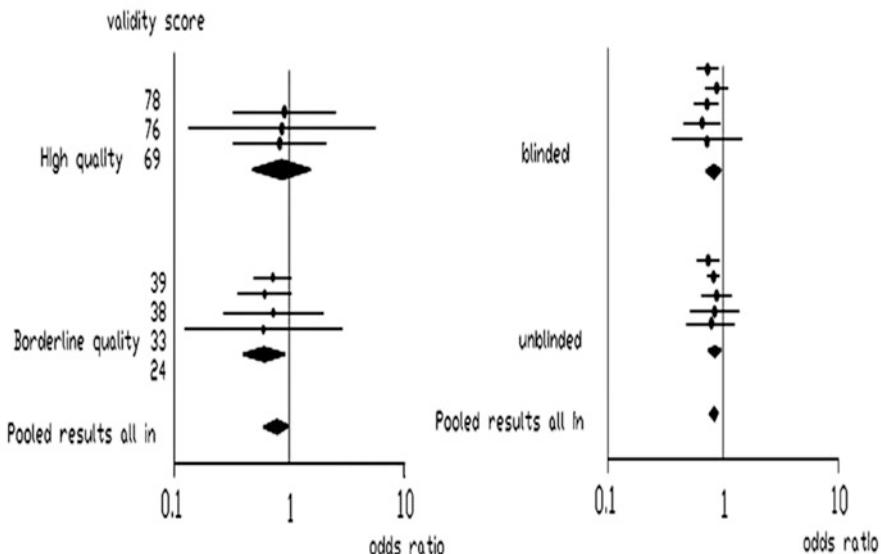
Still other causes for heterogeneity may be involved. As an example, 33 studies of cholesterol and the risk of carcinomas showed that heterogeneity was huge (Khan et al., Arch Int Med. 1996; 156: 661–6). When the trials were divided according to social class, the effect in the lowest class was 4–5 times those of the middle and upper class, explaining everything about this heterogeneous result.

There is some danger of over-interpretation of heterogeneity. Heterogeneity may occur by chance, and will almost certainly be found with large meta-analyses involving many and large studies. This is particularly an important possibility, when no clinical explanation is found, or when the heterogeneity is clinically irrelevant. Also, we should warn, that a great deal of uniformity among the results of independently performed studies is not necessarily good. It can indicate consistency-in-bias rather than consistency-in-real-effects as suggested by Riegelman (Studying a study & testing a test, Lippincott Williams & Wilkins, Philadelphia, PA, USA, 2005, pp. 99–115).

Heterogeneity remains a major issue of meta-analyses, and a core subject in virtually all modern analyses methods. The current edition will review many of them (Chaps. 4, 9, 12, 15, 24, and 25).

## 2.15 Pitfalls, Lack of Sensitivity

Sensitivity or robustness of a meta-analysis is one last aspect to be addressed. When talking of strict inclusion criteria (Chap. 1), we discussed studies with lower levels of validity. It may be worthwhile not to completely reject the studies with lower methodology. They can be worthwhile for assessing sensitivity.



The left graph of the above figure gives an example of how the pooled data of three high-quality-studies provide a smaller result, than do four studies-of-borderline-quality. The summary result is mainly determined by the borderline-quality-studies. When studies are ordered according to their being blinded as shown in the right graph, differences may be large or not. In studies using objective variables, for example blood pressures or heart rates, blinding is not as important as it is in studies using subjective variables (pain scores etc). In this particular example differences were negligible. When examining the influence of various inclusion criteria on the overall odds ratios, we have to conclude that the criteria themselves are an important factor in determining the summary result. In that case the meta-analysis lacks robustness. Interpretation has to be cautious, and pooling may have to be left out altogether. Just leaving out trials at this stage of the meta-analysis is inappropriate either, because it would introduce bias similar to publication-bias or bias-introduced-by-not-complying-with-the-intention-to-treat-principle.

## 2.16 New Developments

Chapter 5 will give examples of data-analysis using the MetaXL statistical software program of Excel. However, many more software programs for the analysis of meta-data are available, for example, by SAS, the Cochrane Revman, S-plus, StatsDirect, StatXact, True Epistat. Most of these programs are expensive, but common procedures are available through Microsoft's Excel and in Excel-add-ins, while many websites offer online statistical analyses for free, including BUGS and R. Leandro's software program (Meta-analysis in medical research. Br Med J books, London UK, 2005) visualizes heterogeneity directly from a computer graph based on Galbraith plots.

In the past few years, many new statistical meta-analysis methods have been developed, and all of them will be duly reviewed in the forthcoming chapters.

Chapter 6 will show that both crossover and parallel-group double blind trials can be analyzed together, although, again, random effect tests should be used to account for the difference in study designs.

Also, this chapter will demonstrate, that with large meta-analyses of randomized controlled trials, currently often called the pinnacle of evidence based research, pitfalls are relatively small, for example smaller than 5%, and that they, therefore, need not always be tested.

Chapter 7 will show, that observational studies observational case -control and cohort studies can be simultaneously included in a meta-analysis, sometimes called meta-epidemiological meta-analyses. Random effect tests should assess the difference in study designs.

Chapter 7 will also show that observational plus randomized studies can be included and that the difference in design can be used for sensitivity analysis.

Chapter 9 will show, that in recent years the method of meta-regression brought new insights. For example, group-level instead of patient-level analyses easily fail to detect heterogeneities between individual patients, otherwise, called ecological biases.

Chapter 10 will show, that meta-analysis is also relevant for pooling the performance of diagnostic tests.

Odds ratios are beautiful, but, without an exact confidence interval, they cannot estimate the magnitude of the populations they have been obtained from. Tetrachoric correlation coefficients are helpful for the purpose (Chap. 20).

With studies heterogeneous due to obviously different populations, contrast coefficients confidence intervals, generally, better fit meta-analysis models, than Satterthwaite confidence intervals do (Chap. 22).

Chapter 25 will address the spin-off of meta-analysis methodologies for other forms of clinical data analysis. Forest plots are, for examples, used for the assessment of interaction and confounding on the outcome.

Chapter 26 will address novel methodologies, for example, the meta-analysis of ANOVAs (analyses of variance), and of data mining data sets.

Agenda-driven bias (cherry-picking and ignoring of studies) and the SPIN phenomenon (specific reporting strategies for bolstering your research and/or finances) are also spin-offs first recognized by meta-analysts (Chap. 26).

Meta-analyses were ‘invented’ in the early 70s by psychologists, but pooling study results extends back to the early 1900s by statisticians such as Karl Pearson, and Ronald Fisher. In the first years pooling of the data was often impossible due to heterogeneity of the studies. However, after 1995 trials became more homogeneous, due to regulations like standardized protocols and many more requirements (Understanding clinical data analysis, learning statistical principles from published clinical research, Chap. 2 Randomized and observational research, Springer Heidelberg Germany, 2016, from the same authors). In the late 90s several publications concluded that meta-analyses did not accurately predict treatment and adverse effects. The pitfalls were held responsible. Initiatives against them include (1) the Consolidated-Standards-of-Reporting-Trials-Movement (CONSORT), (2) the Unpublished-Paper-Amnesty-Movement of the English journals, and (3) the World Association of Medical Editors’ initiative to standardize the peer review system. Guidelines/checklists for reporting meta-analyses were published like QUOROM (Quality of Reporting of Meta-analyses) and MOOSE (Meta-analysis Of Observational Studies in Epidemiology).

## 2.17 Conclusions

Meta-analysis is important in clinical research, because it establishes whether scientific findings are consistent, and can be generalized across populations. The statistical analysis consists of the computation of weighted averages of study characteristics and their standard errors. Common pitfalls of data-analysis are

- (1) publication bias,
- (2) heterogeneity,
- (3) lack of robustness.

New developments in the statistical analysis include

- (1) new software easy to use,
- (2) new arithmetical methods that facilitate the assessment of heterogeneity and comparability of studies,
- (3) a current trend towards more extensive data reporting including multiple subgroup and interaction analyses.

Meta-analyses are governed by the traditional rules for scientific research, and the pitfalls are, particularly, relevant to hypothesis-driven meta-analyses, but less so to current working papers with emphasis on entire data coverage.

## Reference

More background, theoretical and mathematical information of meta-analyses is given in Statistics applied to clinical studies 5th edition, Chaps. 32–34 and 48, Springer Heidelberg Germany, 2012, from the same authors.

# **Chapter 3**

## **Meta-analysis and the Scientific Method**

### **Scientific Rigor and Scientific Method, Two Different Things**

**Abstract** Scientific rigor is the basis of scientific research, and consists of four rules, a prior hypothesis, a thorough data search, strict inclusion criteria, and a thorough data analysis. The prior hypothesis is often called the scientific method, and includes a scientific question, a hypothesis, and a test against control observations. Examples are given.

#### **3.1 Introduction**

In the Chap. 1 the term “scientific rigor” was repeatedly used, a term consistent of four stiff rules of scientific research,

- (1) clearly defined prior hypothesis,
- (2) thorough search,
- (3) strict inclusion criteria, and
- (4) uniform data analysis.

The current chapter focuses on the first rule, a clearly defined prior hypothesis. This is probably the most important one of the four, and it is otherwise often called the scientific method. The condensed version of “the scientific method” is as follows:

reformulate your scientific question into a hypothesis and try and test this hypothesis against control observations.

The scientific method is routinely used in randomized controlled trials, but, otherwise it is not the basis of all kinds of scientific research. The scientific method is not usually applied with observational research. The scientific method is believed to be form of scientific research that is least biased of all forms of scientific research. The daily life of clinical professionals largely consists of routine, with little need for discussion. However, there are questions, that they, simply, do not know the answer to. Some will look for the opinions of their colleagues or the experts in the field. Others will try and find a way out by guessing, what might be the best solution. The benefit of the doubt doctrine (Ordronaux, The jurisprudence of medicine in relation to the law of contracts, and evidence, Lawbook Exchange, 1869) is, often, used as a justification for unproven treatment decisions, and, if

things went wrong, another justification is the expression: clinical medicine is an error-ridden activity (Paget, Unity of mistakes, a phenomenological interpretation of medical work, Comtemp Sociol 1990; 19: 118–9). So far, few physicians routinely follow a different approach, the scientific method.

In clinical settings, this approach is not impossible, but, rarely, applied by physicians, despite their lengthy education in evidence based medicine, which is almost entirely based on the scientific method. One thousand years ago the above-mentioned Ibn Alhazam (965–1040) from Iraq argued about the methods of formulating hypotheses, and, subsequently, testing them. He was influenced by Aristotle and Euclides, from Greece, (300 years BC). Ibn Alhazam on his turn influenced many of his successors, like Isaac Newton (1643–1727), at the beginning of the seventeenth century, from Oxford University UK, a mathematician, famously reluctant to publish his scientific work. His rules of the scientific method were published in a postmortem publication entitled “Study of Natural Philosophy”. They are now entitled the Newton’s rules of the scientific method, and listed in the Oxford English Dictionary, and today routinely used. They are defined, as a method, or, rather, a set of methods, consisting of:

1. a systematic and thorough observation, including measurements,
2. the formulation of a hypothesis regarding the observation,
3. a prospective experiment, and test of the data obtained from the experiment,
4. and, finally, a formal conclusion, and, sometimes, modification of the above hypothesis.

The Cochrane Collaborators advocate the performance of meta-analyses of multiple controlled clinical trials in order to improve the statistical power of the evidence as given by a single controlled clinical trial. The Cochrane group, somewhat, suffers from typically English lack of modesty, and decided, on top of all, to call the result of Cochrane-supported meta-analyses “the pinnacle of scientific knowledge” (Sackett et al, Evidence based medicine 2nd edition, Churchill Livingstone London UK, 2000). However, the Cochrane rules are entirely based on explorative methods using systematic synthesis and combination of results of multiple studies, rather than the above scientific method. In the current chapter, we recommend, that, meta-analyses, not only of randomized controlled trials, but also of any kind of clinical research be based on the scientific method. In this way, they should more properly provide mankind with novel scientific knowledge, rather than explorative uncertainties. As examples, we will rewrite eight recently published exploratory meta-analyses from members of our group into “scientific method” versions.

### 3.2 Example 1, the Potassium Meta-analysis of the Chap. 6

Van Bommel et al. published: Potassium treatment for hypertension in patients with high salt intake a meta-analysis, (Int J Clin Pharmacol Ther 2012; 50: 478–82).

1. Scientific question:  
Does potassium decrease blood pressure in patients with high sodium intake.
2. Hypothesis:  
Potassium tablets of patients with high sodium intake do not reduce blood pressure better than does placebo.
3. Null-Hypothesis testing:  
A multiple groups unpaired t-test of two independent populations of hypertensive patients from multiple parallel-group studies is performed.
4. Result:  
Potassium treatment better reduced high blood pressure in high sodium intake populations than did placebo.

### **3.3 Example 2, the Calcium Channel Blocker Meta-analysis of the Chap. 6**

Cleophas et al. published: Efficacy and safety of second generation dihydropyridine calcium channel blockers in heart failure meta-analysis, Am J Cardiol 2001; 88: 487–90.

1. Scientific question:  
Does calcium channel blockade improve cardiac failure.
2. Hypothesis:  
Calcium channel blockade does not improve cardiac failure.
3. Null-Hypothesis testing:  
A multiple groups unpaired t-test of patients with cardiac failure on calcium channel blocker or placebo from multiple parallel-group studies is performed.
4. Result:  
Calcium channel blockade better improved cardiac index, ejection fraction, and exercise tolerance than did placebo.

### **3.4 Example 3, the Large Randomized Trials Meta-analyses of the Chap. 6**

Zwinderman and Cleophas published: Large meta-analyses of randomized controlled trials need not necessarily be tested for biases, Statistic Sessions, Eudipharm Lyon, December 2000.

1. Scientific question:  
Is it necessary to assess meta-analyses of large randomized trials for the traditional pitfalls of meta-analyses.

**2. Hypothesis:**

The differences between the outcomes adjusted and unadjusted for pitfalls are less than 5%, which is here named the null-hypothesis of no difference.

**3. Null-Hypothesis testing:**

Multiple groups unpaired t-tests comparing the results adjusted and unadjusted for the pitfalls were adjusted for Bonferroni inequality, and showed that differences in results were consistently <5%, and so the null-hypothesis of no difference could never be rejected.

**4. Result:**

It is not necessary to assess meta-analyses of large randomized trials for traditional pitfalls of meta-analyses.

### **3.5 Example 4, the Diabetes and Heart Failure Meta-analysis of the Chap. 7**

Kamalesh et al. published: Heart Failure due to systolic dysfunction and mortality in diabetes: pooled analysis of 39,505 subjects, J Card Fail 2009; 15: 3005–9.

**1. Scientific question:**

Does diabetes mellitus increase the risk of heart failure

**2. Hypothesis:**

Diabetes mellitus does not increase the risk of heart failure.

**3. Null-Hypothesis testing:**

A multiple groups unpaired chi-square test tests the odds ratios of heart failure in patients with and without diabetes mellitus.

**4. Result:**

The difference was very significantly so, the null hypothesis was rejected.

### **3.6 Example 5, the Adverse Drug Effect Admissions and the Type of Research Group Meta-analysis of the Chap. 8**

Atiqi et al. published: Prevalence of iatrogenic admissions to the departments of medicine/cardiology/pulmonology in a 1250 beds general hospital, Int J Clin Pharmacol Ther 2009; 47: 549–56.

**1. Scientific question:**

Is the type of research group a determinant of the numbers of adverse drug effect admissions to hospital.

**2. Hypothesis:**

The type of research group is not a determinant

## 3. Null-Hypothesis testing:

In a linear regression controlled for age and study size it is assessed whether pharmacists report adverse drug effects more often than internists do.

## 4. Result:

The difference was very significantly so, the null hypothesis was rejected.

### 3.7 Example 6, the Coronary Events and Collaterals Meta-analysis of the Chap. 8

Akin et al. published: Effect of collaterals on deaths and re-infarctions in patients with coronary artery disease a meta-analysis, Neth Heart J 2012; 21: 146–51.

## 1. Scientific question:

Do coronary collaterals protect against coronary events.

## 2. Hypothesis:

Coronary collaterals do not protect.

## 3. Null-Hypothesis testing:

A multiple groups unpaired chi-square test tests the odds ratios of coronary events in patients with collaterals versus those without.

## 4. Result:

Patients with coronary collaterals are less at risk of myocardial infarction than those without.

### 3.8 Example 7, the Diagnostic Meta-analysis of Metastatic Lymph Node Imaging of the Chap. 10

Cleophas et al. published: Meta-analysis of qualitative diagnostic tests, Chap. 48, in: Statistics applied to clinical studies 5th edition, pp. 527–34, Springer Heidelberg Germany, from the same authors.

## 1. Scientific question:

Is Magnetic resonance imaging better than computerized tomography or lymphangiography for diagnosing lymph node metastases.

## 2. Hypothesis:

The three above methods are not significantly different from one another.

## 3. Null-Hypothesis testing:

Linear regression of invert logs of diagnostic odds ratios compared the magnitudes of the b-values (regression coefficients).

## 4. Result:

Magnetics resonance imaging performed significantly better at  $p < 0.0001$ .

### 3.9 Example 8, the Homocysteine and Cardiac Risk Meta-analysis of the Chap. 11

Cleophas et al. published: Homocysteine a risk factor for coronary artery disease or not, Am J Cardiol 2000; 86: 1005–9.

1. Scientific question:

Does homocysteine increase cardiac risks.

2. Hypothesis:

Homocysteine patients carry no increased cardiac risks as compared to control patients.

3. Null-Hypothesis testing:

A multiple groups unpaired chi-square test tests the odds ratios of cardiac risk in patients with and without homocysteine levels.

4. Result:

The above test was very significant, but the studies were very heterogeneous, both according to the fixed and the random effect tests.

### 3.10 Conclusions

The term scientific rigor is a term consisting of four rigorous, meaning stiff, rules of scientific research,

- (1) clearly defined prior hypothesis,
- (2) thorough search,
- (3) strict inclusion criteria, and
- (4) uniform data analysis.

The current chapter focused on the first rule, a clearly defined prior hypothesis. This is probably the most important one of the four, and it is otherwise often called “the scientific method”. A short description of “the scientific method” is as follows:

reformulate your scientific question into a hypothesis and try and test this hypothesis against control observations.

The scientific method is routinely used in randomized controlled trials, but, otherwise it is not the basis of all kinds of scientific research. The scientific method is not usually applied with observational research. The scientific method is believed to be form of scientific research that is least biased of all forms of scientific research. In clinical settings, this “scientific method approach” is rarely applied by medical professionals. A more extensive description of "the scientific methods was described by Newton:

1. a systematic and thorough observation, including measurements,
2. the formulation of a hypothesis regarding the observation,

3. a prospective experiment, and test of the data obtained from the experiment,
4. and, finally, a formal conclusion, and, sometimes, modification of the above hypothesis.

The Cochrane Collaborators advocate the performance of meta-analyses of multiple controlled clinical trials in order to improve the statistical power of the evidence as given by a single controlled clinical trial, and decided to call the result of Cochrane-supported meta-analyses “the pinnacle of scientific knowledge” (Sackett et al., Evidence based medicine 2nd edition, Churchill Livingstone London UK, 2000). However, The Cochrane rules are entirely based on explorative methods using systematic synthesis and combination of results of multiple studies, rather than the above scientific method. In the current chapter, we conclude, that, meta-analyses, not only of randomized controlled trials, but also of any kind of clinical research be based on the scientific method. In this way, they should more properly provide mankind with novel scientific knowledge, rather than explorative uncertainties.

We should add, that “the scientific method” is very much determined by randomness and probabilities. Traditionally, probabilities were defined as prior knowledge, either the word of God or any other prior likelihood, often based on individual experiences followed by logical interpretation. Statistics with traditional probabilities is called Bayesian statistics. Some 100 years ago another type of probability was invented: physical probability, also called objective probability. And it became widely accepted. John Venn (Cambridge UK, 1834–1923), and Ronald Fisher (Cambridge UK and Adelaide Australia, 1890–1960) rejected the Bayesian probability and started to propagate physical probability. A group of statisticians, called the frequentists, agreed with the two, and frequentism became the most important school in statistics, as the scientific study of probabilities. Ronald Fisher invented and gave the name to the famous f-test, the basic test for frequentists, namely analysis of variance. In the 70s meta-analyses were performed by psychologists. They were systematic reviews without data pooling. Then the frequentists took over, and pooling of outcome data started. Meta-analysis, as we know it today, covers a very much frequentists’ school of analytical methodologies. Frequentists rely, more than anything else, on the scientific method with its null-hypotheses. That is why the scientific method will be so important, when you are involved in meta-analyses.

## References

- Akin S et al (2012) Effect of collaterals on deaths and re-infarctions in patients with coronary artery disease a meta-analysis. Neth Hear J 21:146–151
- Atiqi R et al (2009) Prevalence of iatrogenic admissions to the departments of medicine/cardiology/pulmonology in a 1250 beds general hospital. Int J Clin Pharmacol Ther 47:549–556
- Cleophas TJ et al (2000) Homocysteine a risk factor for coronary artery disease or not. Am J Cardiol 86:1005–1009

- Cleophas TJ et al (2001) Efficacy and safety of second generation dihydropyridine calcium channel blockers in heart failure meta-analysis. *Am J Cardiol* 88:487–490
- Cleophas TJ et al (2014) Meta-analysis of qualitative diagnostic tests, Chap. 48. In: Statistics applied to clinical studies, 5th ed, Springer, Heidelberg, pp 527–534, from the same authors
- Kamalesh M et al (2009) Heart Failure due to systolic dysfunction and mortality in diabetes: pooled analysis of 39,505 subjects. *J Card Fail* 15:3005–3009
- Van Bommel E et al (2012) Potassium treatment for hypertension in patients with high salt intake a meta-analysis. *Int J Clin Pharmacol Ther* 50:478–482
- Zwinderman AH and Cleophas TJ (2000) Large meta-analyses of randomized controlled trials need not necessarily be tested for biases, Statistic Sessions, Eudipharm Lyon (France), December 2000

# **Chapter 4**

## **Meta-analysis and Random Effect Analysis**

### **Old and New Style Random Effect Analysis**

**Abstract** Heterogeneity is probably the largest pitfall of meta-analyses. A random effect analytic model assumes, that heterogeneity is due to some unexpected subgroup effect rather than a residual effect, and uses a separate random variable for the purpose. Examples of fixed and random effect analyses are given both from binary and continuous outcome meta-analysis data. Within a single study heterogeneity may very well be residual, but between the overall effects of studies within a meta-analysis this is virtually never so, and it is virtually always caused by some subgroup effect. Therefore, fixed effect heterogeneity tests are slightly inappropriate. However, random effect heterogeneity tests lack power. Novel methodologies are being developed, and will be reviewed in this and many subsequent chapters of this edition.

#### **4.1 Introduction**

Heterogeneity in a clinical trial means, that the differences in the results between the subjects are larger than could happen by chance. Heterogeneity is commonly assumed to be due to an inherent variability in biological processes, sometimes called a residual effect, rather than some hidden subgroup property, and this is named a fixed effect. It will happen again and again, if you repeat the research. If, however, we have reasons to believe, that certain patients, due to co-morbidity, co-medication, age or other factors, will respond differently from others, then the spread in the data will be caused, not only by the residual effect, but also by between patient differences due to some subgroup property. It will probably not happen again, if you repeat the research at another time or place. With the fixed effect model the treatment differences are tested against the residual error, and often the standard error is used for the purpose. With the random effect models the treatment effects is influenced not only by the residual effect, but also by an unexpected, otherwise called random, factor, and, so, the treatment should not only be tested against the residual effect, but, in addition, against the random factor. Random effect analyses are sometimes called advanced analysis of variance or mixed effects models (Chap. 56, Statistics applied to clinical studies 5th edition, 2012, Springer Heidelberg Germany, from the same authors), and they are a very

interesting class of models, although even a partial understanding of them is fairly difficult to achieve.

In meta-analyses heterogeneity may be caused not only by differences between individuals but also by differences between entire studies as included. Within a study the heterogeneity may very well be residual, but between the overall effects of the studies this is virtually never so, and it is virtually always caused by some random subgroup effect. It is, therefore, silly, and, actually, pretty inappropriate to perform fixed effect analyses for heterogeneity in meta-analyses. This point has been emphasized in recent methodological studies (Dersimonian and Kacker, Contemporary Clinical Trials, doi 10.1016/j.cct.2006, Hodges, The American Statistician 2010; 64: 325–34. Anonymous, Meta-Analysis, Wikipedia April 2016). The fixed effect heterogeneity models will demonstrate heterogeneity at a higher level of significance, than the random effect models do, and the common approach is to perform a fixed effect model first, and, if statistically significant, a random effect model second. And so, if the fixed effects tests is not significant, then the random effect test will certainly not be significant, and you need not do the random effect tests anymore. The reasoning for a random effect analysis is slightly inappropriate here, because, rather than the type of heterogeneity, the magnitude of the p-value is used as the main reason for the choice of the ultimate analysis method. Hodges and Clayton (Random effect old and new, [www.semanticscholar.org](http://www.semanticscholar.org), February 2, 2011) call the latter type of random effect analysis new-style random effect analysis. It has nothing to do with random effect anymore, but, instead, with the justification for using a test that best shrinks the level of heterogeneity, because the use of random effect models will lead to wider confidence intervals and a lower chance to call a difference statistically significant. Another disadvantage of the random-effect analysis is (Berlin et al. stat Med 1989; 8: 141–151), that small and large studies are given almost similar weights.

More robust methods for testing heterogeneity than the traditional random effect models are, obviously, welcome, and, particularly, the inverted variance heterogeneity (IVhet) method of Barendrecht (Doi and Barendregt, Advances in the meta-analysis of heterogeneous clinical trials, Comtemp Clin Trials 2015 doi:101016) and the quality effect models (MetaXL User Guide, 2016, [www.epigear.com](http://www.epigear.com), pp. 36–40) are promising replacements. They will be reviewed and assessed in the next chapter.

The traditional fixed effect model for testing heterogeneity of binary outcome studies is the Cochran Q test, otherwise called Q-statistic (Chap. 2 gives the details). Complementary to the Q-statistic, the amount of heterogeneity between studies is often quantified with the  $I^2$ -statistic (Higgins and Thompson, Stat Med 2002; 21: 1539–58),

$$I^2 = 100\% * [Q - (k - 1)]/Q$$

which is interpreted as the proportion of total variation in study estimates due to heterogeneity rather than sampling error. Fifty % is often used as a cut-off for heterogeneity. What is the place of the  $I^2$  test. For binary data it is, actually, equal to

the fixed effect test. In the current chapter we will review random effect models for heterogeneity both for studies with binary and for those with continuous outcome data.

## 4.2 Visualizing Heterogeneity

In clinical research similar studies often have different results. This may be due to differences in patient-characteristics and trial-quality-characteristics such as the use of blinding, randomization, and placebo-controls. Three-dimensional scatter plots are sometimes able to identify mechanisms responsible.

Variables	1 % ADEs	2 study size	3 age	4 investigator type
21,00	106	1	1	
14,40	578	1	1	
30,40	240	1	1	
6,10	671	0	0	
12,00	681	0	0	
3,40	28411	1	0	
6,60	347	0	0	
3,30	8601	0	0	
4,90	915	0	0	
9,60	156	0	0	
6,50	4093	0	0	
6,50	18820	0	0	
4,10	6383	0	0	
4,30	2933	0	0	
3,50	480	0	0	
4,30	19070	1	0	
12,60	2169	1	0	
33,20	2261	0	1	
5,60	12793	0	0	
5,10	355	0	0	

ADEs = adverse drug effects

age 0 = young, 1 = elderly

investigator type, 0 = pharmacists, 1 = clinicians

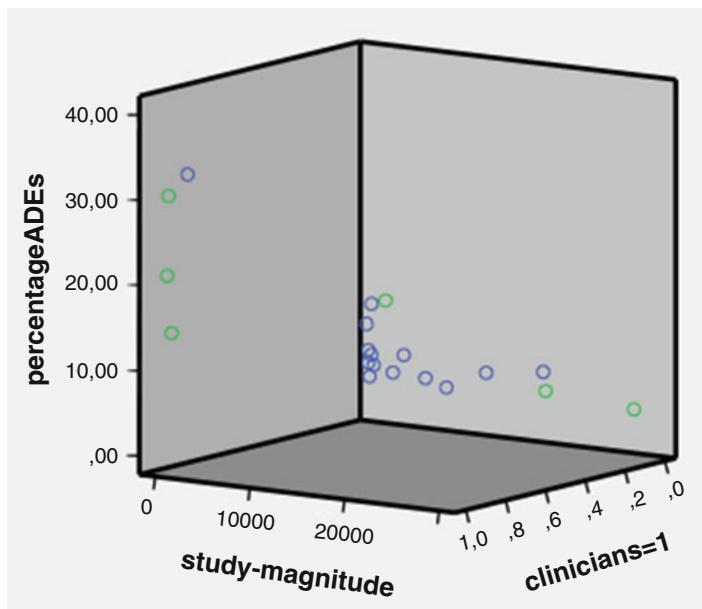
In the above 20 studies the percentage of admissions to hospital due to adverse drug effects were assessed. The studies were very heterogeneous, because the percentages admissions due to adverse drug effects varied from 3.3 to 33.2. In order to identify possible mechanisms responsible, a scatter plot was first drawn.

The data file is in extras.springer.com and is entitled “heterogeneity”. Start by opening the data file in SPSS statistical software.

### Command

click Graphs....click Legacy Dialogs....click Scatter/Dot....click 3-D Scatter....click Define....Y-Axis: enter percentage (ADEs)....X Axis: enter study-magnitude....Z Axis: enter clinicians =1....Set Markers by: enter elderly =1....click OK.

The underneath figure is displayed, and it gives a three-dimensional graph of the outcome (% adverse drug effects) versus study size versus investigator type (1 = clinician, 0 = pharmacist). A fourth dimension is obtained by coloring the circles (green = elderly, blue = young). Small studies tended to have larger results. Also clinician studies (clinicians =1) tended to have larger results, while studies in elderly had both large and small effects.



In order to test whether the observed trends were statistically significant, linear regressions can be performed.

### 4.3 Binary Outcome Data, Fixed Effect Analysis

The underneath data of a coronary collaterals meta-analysis is a meta-analysis of open evaluation studies, and show the results of 7 studies assessing chance of death and infarction in patients with coronary collaterals compared to that in patients without.

	Odds Collaterals	odds no collaterals	n	odds ratio	95% ci	z-value	p
1.Monteiro 2003	6/29	11/24	70	0.45	0.15-1.40	-1.38	1.69
2.Nathou 2006	3/173	20/365	561	0.32	0.09-1.08	-1.84	0.066
3.Meier 2007	36/190	197/389	812	0.37	0.25-0.56	-4.87	0.0001
4.Sorajja 2007	7/112	15/184	318	0.77	0.30-1.94	-0.56	0.576
5.Regieli 2009	7/254	16/600	879	1.03	0.42-2.54	+0.07	0.944
6.Desch 2010	5/64	34/132	235	0.30	0.11-0.81	-2.38	0.018
7.Steg 2010	246/1676	42/209	2173	0.73	0.51-1.04	-1.72	0.085

In order to meta-analyze these data, the following calculations are required.  
OR = odds ratio, lnOR = the natural logarithm of the odds ratio, var. = variance.

	OR	lnOR	var	1/var	lnOR/var	(lnOR) <sup>2</sup> /var
1.Monteiro 2003	0.45	-0.795	0.3337	2.997	-2.382	1.894
2.Nathou 2006	0.32	-1.150	0.3919	2.882	-2.935	3.375
3.Meier 2007	0.37	-0.983	0.04069	24.576	-24.158	23.748
4.Sorajja 2007	0.77	-0.266	0.2239	4.466	-1.188	0.3160
5.Regieli 2009	1.03	-1.194	0.2526	3.959	-4.727	5.644
6.Desch 2010	0.30	0.032	0.2110	4.739	0.152	0.005
7.Steg 2010	0.73	-0.314	0.0333	30.03	9.429	2.961
				73.319	-44.667	37.943

The pooled odds ratio is calculated from:

$$\text{antiln of } (-44.667/73.319) = 0.54.$$

The chi-square value for the above pooled data

$$\begin{aligned} &= (-44.667)^2/73.319 \\ &= 27.2117. \end{aligned}$$

According to the chi-square table, for a chi-square value  $> 10.827$  and 1 degree of freedom, this would correspond to a p-value  $= <0.001$ . This would mean that overall the patients with collateral coronary arteries have far less chance of death or infarction.

The above data will now be assessed for heterogeneity. Heterogeneity of this meta-analysis is tested by the fixed effect model.

$$\begin{aligned} \text{Heterogeneity chi-square value} &= 37.943 - 27.2117 \\ &= 10.7317 \end{aligned}$$

With 6 degrees of freedom a chi-square value  $> 10.645$  would mean a p-value  
 $= 0.05 < p < 0.10$ .

Although the meta-analysis shows a significantly lower risk in patients with collaterals than in those without, this result has a limited meaning, since there is a trend to heterogeneity in these studies. For heterogeneity testing it is tested, whether the differences between the results of the separate trials are greater than compatible with the play of chance. Additional tests for heterogeneity testing are available (Cleophas and Zwinderman, Meta-analysis. In: Statistics Applied to Clinical Studies, Springer New York, 2012, 5th edition, pp. 365–388). When there is heterogeneity in a meta-analysis, a careful investigation of its potential cause is often more important than a lot of additional statistical tests.

## 4.4 Binary Outcome Data, Random Effect Analysis

A fixed effect model assumes, that the amount of heterogeneity will be unchanged any time the study is repeated. However, a different approach is, to assume, that the heterogeneity is caused by some unexpected subgroup property, and that it will not occur the next time you will perform the same study. If we have reasons to believe, that the patients of one study, due to unmeasured co-morbidity, co-medication, age or other factors will respond differently from others, then the heterogeneity in the studies will be caused not only by the residual effect, but also by between study differences, due to study specific properties. It may be safe, to routinely treat any heterogeneity as a random effect, unless there are good arguments not to do so. Random effect meta-analysis models require a statistical approach different from that of fixed effect models. Particularly, the Dersimonian Laird model is frequently applied, and provides some shrinkage of the weighting factors and widening of the variances of the individual studies in a meta-analysis.

	OR y	lnOR	var	1/var w	lnOR/var w.y	(lnOR) <sup>2</sup> /var w.y <sup>2</sup>
1.Monteiro 2003	0.45	-0.795	0.3337	2.997	-2.382	1.894
2.Nathou 2006	0.32	-1.150	0.3919	2.882	-2.935	3.375
3.Meier 2007	0.37	-0.983	0.04069	24.576	-24.158	23.748
4.Sorajja 2007	0.77	-0.266	0.2239	4.466	-1.188	0.3160
5.Regieli 2009	1.03	-1.194	0.2526	3.959	-4.727	5.644
6.Desch 2010	0.30	0.032	0.2110	4.739	0.152	0.005
7.Stege 2010	0.73	-0.314	0.0333	30.03	9.429	2.961
				73.319	-44.667	37.943

The Dersimonian-Laird method uses, per study included, a weighting factor slightly different from the fixed effect weighting factor, namely  $w^*$ :

$$w^* = 1/(D + 1/w) \text{ (if } w \text{ is vary large, then } w^* \text{ turns into } 1/D)$$

The term D:

$$\begin{aligned} &= \frac{[\Sigma (\ln\text{OR})^2/\text{var} - \text{chi-square value for pooled data}] - (7-1) (\Sigma (1/\text{var}))}{\Sigma (1/\text{var}) - \Sigma (1/\text{var})^2} \\ &= \frac{10.7317 - 439.914}{73.319 - 1581.148} \\ &= 0.2846 \end{aligned}$$

$$w^* \text{ for study 1} = 1 / (D + 1/w) = 1 / (0.2846 + (1/2.997)) = 1.614 \quad w^*\ln\text{OR} = -1.28$$

$$w^* \text{ for study 2} = 1 / (D + 1/w) = 1 / (0.2846 + (1/2.882)) = 1.583 \quad w^*\ln\text{OR} = -1.82$$

$$w^* \text{ for study 3} = 1 / (D + 1/w) = 1 / (0.2846 + (1/24.576)) = 3.074 \quad w^*\ln\text{OR} = -3.02$$

$$w^* \text{ for study 4} = 1 / (D + 1/w) = 1 / (0.2846 + (1/4.466)) = 1.9665 \quad w^*\ln\text{OR} = -0.52$$

$$w^* \text{ for study 5} = 1 / (D + 1/w) = 1 / (0.2846 + (1/3.959)) = 1.8615 \quad w^*\ln\text{OR} = -2.22$$

$$w^* \text{ for study 6} = 1 / (D + 1/w) = 1 / (0.2846 + (1/4.739)) = 2.0178 \quad w^*\ln\text{OR} = 0.065$$

$$w^* \text{ for study 7} = 1 / (D + 1/w) = 1 / (0.2846 + (1/30.030)) = 3.1456 \quad w^*\ln\text{OR} = -0.99$$

$$\Sigma = 15.26 \quad \Sigma = -9.785$$

$$\text{OR} = e^{-(-9.785/15.26)} = 0.527 \text{ (^ announces superscript term)}$$

95% confidence interval of this OR

$$= e^{(\ln\text{OR} \pm 1.96/\sqrt{\Sigma w^*})}$$

$$= e^{(-0.64 \pm 1.96/\sqrt{15.26})}$$

= between 0.32 and 1.15.

The pooled odds ratio of the random effect model is almost the same as that of the fixed effect model, 0.527 and 0.54. The confidence interval of the fixed effect

model is between 0.41 and 0.73. The random effect confidence interval is much wider and statistically insignificant ( $t = -0.64/0.51$ , and, thus, is much larger than  $-2$ , the traditional 0.05 level of significance).

This would mean that the random effect model produced a similar pooled odds ratio, with the chance of death and infarction without collaterals about twice as high as that with collaterals. The fixed effect heterogeneity test produced a tendency to heterogeneity, while the random effect test produced an insignificant test for heterogeneity.

Generally, the random effect model is considered required, if the fixed effect model produced heterogeneity. If confirmed, heterogeneity will be assumed to be in the study, and pooling is no further warranted, and the study should be analyzed as a systematic review, rather than a meta-analysis.

However, the random effect model is recently being scrutinized. For example, it can be observed in the above two tables, that, in the fixed effect model, with large variances the weighting factors are much larger than they are with small variances. In contrast, in the random effect model, with large variances, the weighting factors are just a little bit larger, than they are with small variances. The smoothing process, thus, seems to be selective, and to mainly affect the studies with small variances, a pretty silly phenomenon. It was the reason for criticisms of the Dersimonian and Laird methodology (Hodges et al. Random effect old and new, CRC Press, 2013; 285–302), and some authors even recommended, that meta-analyses are no good at all, because of heterogeneity due to interstudy subjective considerations and assessments (Stegenga, Is meta-analysis the platinum standard of evidence, Studies in history and philosophy of biological and biomedical sciences 2011; doi: 10.1016). See also Eysenck (Systematic reviews: meta-analysis and its problems, BMJ 1994; 309: 780–92) and Hill (The environment and disease: association or causation, Proc Roy Soc Med 1965; 58: 295–300) for discussions on meta-analysis problems and common sense reasonings on heterogeneous data. Hill wrote guidelines for meta-analyses *avant la lettre*, giving a plurality of common sense reasonings for assessing heterogeneity between different studies including strength of statistical associations, consistency of results, specific effects, temporality, factors like dose-response patterns, plausibility arguments, coherence of the studies with other relevant knowledge etc. Statements in any meta-analysis are not often judgments supported by values, but, rather, just assumptions. Nonetheless, if one accepts all of the flaws, it still is a fascinating explorative post-hoc methodology, in support of progress of scientific knowledge.

## 4.5 Continuous Outcome Data, Fixed Effect Analysis

The underneath data are from the potassium meta-analysis of the Chap. 6, a meta-analysis of double blind placebo controlled trials. The meta-analysis assessed the difference in systolic blood pressures (mm Hg) between patients treated with potassium and those with placebo.

Diff = difference in systolic blood pressure between patients on potassium and placebo, var. = variance = (standard error)<sup>2</sup>

	N	diff	standard (systolic) error	1/var	diff/var	diff <sup>2</sup> /var
1.McGregor 1982	23	-7.0	3.1	0.104	-0.728	5.096
2.Siani 1987	37	-14.0	4.0	0.063	-0.875	12.348
3.Svetkey 1987	101	-6.4	1.9	0.272	-1.773	11.346
4.Krishna 1989	10	-5.5	3.8	0.069	-0.380	2.087
5.Obel 1989	48	-41.0	2.6	0.148	-6.065	248.788
6.Patki 1990	37	-12.1	2.6	0.148	-1.791	21.669
7.Fotherby 1992	18	-10.0	3.8	0.069	-0.693	6.900
8.Brancati 1996	87	-6.9	1.2	0.694	-4.792	33.041
9.Gu 2001	150	-5.0	1.4	0.510	-2.551	12.750
10.Sarkkinen 2011	45	-11.3	4.8	0.043	-0.490	5.091
						+
				2.125	-20.138	359.516

$$\text{Pooled difference} = -20.138/2.125 = -9.48 \text{ mm Hg}$$

$$\text{Chi-square value for pooled data} = (-20.138)^2/2.125 = 206.91$$

According to the chi-square table the p-value for 1 degree of freedom  
= < 0.001.

The above data will now be assessed for heterogeneity. Heterogeneity of this meta-analysis is tested by the fixed effect model.

$$\text{Pooled difference} = -20.138/2.125 = -9.48 \text{ mm Hg}$$

$$\text{Chi-square value for pooled data} = (-20.138)^2/2.125 = 206.91$$

$$\begin{aligned} \text{Heterogeneity chi-square value} &= 359.516 - 206.91 \\ &= 152.6, \end{aligned}$$

With 9 degrees of freedom the p –value  
= < 0.001.

Although the meta-analysis shows a significantly lower systolic blood pressure in patients with potassium treatment than those with placebo, this result has a limited meaning, since the studies are significantly heterogeneous. For heterogeneity testing it is tested, whether there is a greater inequality between the results of the separate trials than compatible with the play of chance. Additional tests for heterogeneity testing are available (Cleophas and Zwinderman, Meta-analysis. In: Statistics Applied to Clinical Studies, Springer New York, 2012, 5th edition, pp. 365–388). However, when there is heterogeneity, a careful investigation of its potential cause is more important than lots of additional statistical tests.

## 4.6 Continuous Outcome Data, Random Effect Analysis

	N	diff	standard (systolic) error	1/var	diff/var	diff <sup>2</sup> /var
1.McGregor 1982	23	-7.0	3.1	0.104	-0.728	5.096
2.Siani 1987	37	-14.0	4.0	0.063	-0.875	12.348
3.Svetkey 1987	101	-6.4	1.9	0.272	-1.773	11.346
4.Krishna 1989	10	-5.5	3.8	0.069	-0.380	2.087
5.Obel 1989	48	-41.0	2.6	0.148	-6.065	248.788
6.Patki 1990	37	-12.1	2.6	0.148	-1.791	21.669
7.Fotherby 1992	18	-10.0	3.8	0.069	-0.693	6.900
8.Brancati 1996	87	-6.9	1.2	0.694	-4.792	33.041
9.Gu 2001	150	-5.0	1.4	0.510	-2.551	12.750
10.Sarkkinen 2011	45	-11.3	4.8	0.043	-0.490	5.091
					+	
	556			2.125	-20.138	359.516

In the example of Sect. 4.5, the fixed effect tests for heterogeneity were statistically significant, and, thus, a random effect analysis needs to be performed in order to assess, whether this effect was also statistically significant, if we assume the heterogeneity to be due to an unexpected subgroup effect in our data, that will not happen again the next time. The random effect test model produces a much smaller test statistic than did the fixed effect model for heterogeneity.

$$\begin{aligned} (\Sigma(\text{diff}^2/\text{var})/\text{N} - \Sigma(\text{diff}/\text{var})^2/\text{N}^2) &= 359.516/556 - (-20.138)^2/(556)^2 \\ &= 0.6466 - 0.0013 \\ &= 0.6453 \end{aligned}$$

$$\begin{aligned} (\Sigma\text{diff}/\text{var})/(\text{N}) &= -20.138/556 \\ &= -0.036 \end{aligned}$$

The variance of the random effect model  $s^*$  is given by

$$\begin{aligned} s^* &= 4k/\text{N} (1 + [(\Sigma\text{diff}/\text{var})/(\text{N})]^2/8) \\ &= 0.07 \times 1.00016 \\ &= 0.07001 \end{aligned}$$

$$\begin{aligned} \text{The chi-square value} &= 10 \times (0.6453)^2/0.07001 \\ &= 59.48. \end{aligned}$$

Now, with 10–1 degrees of freedom the chi-square value of at least 27.88 would be needed to obtain a p-value of 0.001 or less. Our chi-square, however, is a lot smaller than the fixed effect chi-square of 206.91, but, nonetheless, a very significant random effect heterogeneity was in these data.

And so, both the fixed and the random effect heterogeneity tests were statistically significant, and, thus, pooling these data make no sense anymore. The results must be reported as a systematic review, without weighted summary measures. Just like with the binary metadata of Sects 4.4 and 4.5 of this chapter, some smoothing (shrinkage of the variances) is observed, leading in the binary example to a change from tendency to significance to insignificance, and leading in the current example

of continuous data to very high significance to significance at a somewhat lower level.

## 4.7 Conclusions

Fixed effect models assume, that an intervention has only a single true effect, while random effect models assume, that the effect may be different between different studies. With the fixed effect model the treatment differences are tested against the residual error, otherwise called the standard error. With the random effect models the treatment effects may be influenced not only by the residual effect but also by some unexpected, otherwise called random, factor, and so the treatment should no longer be tested against the residual effect, but against the random effect. In meta-analyses heterogeneity may be caused not only by differences between individuals but also by differences between the overall effects of entire studies as included. Within a study the heterogeneity may very well be residual, but between the overall effects of the studies this is virtually never so, and the heterogeneity is virtually always caused by some random subgroup effect. It is, therefore, pretty silly to perform random effect analyses for heterogeneity in meta-analyses. Nonetheless, as the fixed effect heterogeneity models will demonstrate heterogeneity at a lower level of significance than the random effect models, the common approach to first use the fixed effect models, and, if significant, the random effect models second is widely applied despite its silliness. And so, if the fixed effects tests is not significant, then the random effect test will certainly not be significant, and you need not do the random effect tests anyway.

In the next chapter we will address the use of IVhet and quality effect heterogeneity models, as more robust alternatives to the current random effect model.

What is the place of the  $I^2$  test. More information has already been given in the Chap. 2. For binary data, it is, actually, equal to the fixed effect test. In the current chapter we reviewed random effect models for heterogeneity both for studies with binary and for those with continuous outcome data. The examples of this chapter show, that random effect have smaller test statistics than fixed effect models have. This is common and appropriate for statistical models in general: the better and less biased the model, the smaller and less powerful the test statistic.

We should add, that more modern alternatives to the traditional random effect models were recently proposed. Bonett (Meta-analytic interval estimation for standardized and unstandardized mean differences, Psych Meth 2009; 14: 225–38) proposed the so-called changing contrast coefficient model and Gagnier et al. (Syst Rev. 2012; 1: 18) the closely related permutation models. They will be addressed in the Chap. 22, but produce pretty similar results.

## Reference

More background, theoretical and mathematical information of meta-analyses is given in Statistics applied to clinical studies 5th edition, Chaps. 32–34 and 48, Springer Heidelberg Germany, 2012, from the same authors.

# **Chapter 5**

## **Meta-analysis with Statistical Software**

### **Quasi Likelihood Modeling, Adjusted Heterogeneity Without Overdispersion**

**Abstract** Traditional meta-analytic computations are pretty straightforward and do not necessarily need the help of statistical software. Unfortunately, traditional fixed effect analysis is slightly inappropriate, random effect analysis is weak. Analyses with novel methodologies like the quasi-likelihood methodology and the quality effects method are welcome, but they can not be accomplished without a computer. Examples and step by step analyses are given.

#### **5.1 Introduction**

This chapter will give examples of data-analysis using meta-analysis software programs, using free calculators from the internet and the MetaXL program at [www.epigear.com](http://www.epigear.com). However, many more software programs for the analysis of meta-data are available, for example, SAS software, the Cochrane Revman, S-plus, StatsDirect, StatXact, True Epistat and other statistical software programs. Most of these programs are expensive, e.g., the software program Comprehensive Meta-Analysis from Arlington University Virginia has a very user-friendly menu, but will cost you about \$1000 per year per copy. Common procedures are also available through Microsoft's Excel and in Excel-add-ins, while many websites offer online statistical analyses for free, including BUGS and R. Leandro's software program (Meta-analysis in medical research. Br Med J books, London UK, 2005) visualizes heterogeneity directly from a computer graph based on Galbraith plots.

#### **5.2 Using Online Meta-analysis Calculators and MetaXL Free Meta-analysis Software**

Meta-analysis Calculator from [www.healthstrategy.com](http://www.healthstrategy.com) is adequate for pooling and fixed effect heterogeneity test of studies with continuous outcome data. It is convenient and rapid. a data example will be given in the next section.

MetaXL free software for Excel, recently developed by Barendregt, associate professor of epidemiologic modeling at the University of Brisbane, Queensland Australia (Doi and Barendregt, Advances in the meta-analysis of heterogeneous

clinical trials, Comtemp Clin Trials 2015 doi:101,016). When downloaded, it is automatically stored in your computer's excel program. It has many traditional options including pooling options, heterogeneity assessments, and it includes the possibility to draw various types of graphs.

A special point of the MetaXL program is, that, since version 2, it uses a novel procedure for assessing heterogeneity between studies. It is called the invert variance heterogeneity (IVhet) method. It is based on quasi likelihoods. When the fixed effect analysis is significant, a random effect analysis should be performed, but this analysis is virtually always statistically insignificant, because of the wide variances (overdispersion), and therefore, does not mean too much. Quasi likelihoods are like Gaussian likelihood ratios. Likelihood tests are a bit like traditional tests, e.g., z-and t-tests, but they are more sensitive, because they are *exact* tests, comparable, e.g., with Fisher's exact test. More information of the calculus is given in Statistics applied to clinical studies 5th edition, Chap. 4, Log likelihood tests, Springer Heidelberg Germany, 2012, from the same authors. These tests are wonderful, because, with large and small samples, they provide better p-values than the traditional z – and chisquare – tests do.

How does it work. Some features of the data are automatically specified, e.g., data being continuous or not, variability changes with averages, dependent or independent data, medians or means, and the effects on them from the remainder of the data available, etc. From that exercise by the software program a better fit model than that of a traditional likelihood model is obtained and selected. For example, in your data a higher variance in the center of your data is observed by the software program. Therefore, the software program concludes, that the data are (pseudo-) binomial, rather than continuous. It, then, changes the standard deviation (SD) of your mean into the formula appropriate for binomial analysis, with standard deviations computed from binomials like mean x (1-mean) terms. Calculations will be pretty tough, but, for a computer, this is no problem.

### 5.3 Continuous Outcome Data, Online Meta-analysis Calculator

The online meta-analysis calculator entitled Meta-Analysis-Calculator from [www.healthstrategy.com](http://www.healthstrategy.com) (Health Decision Strategies LLC (limited liability company), Princeton, NJ) was used for analyzing a meta-analysis of 8 parallel-group studies in patients with hypertension with mean systolic blood pressure (mm Hg) as outcome measure. The data are underneath.

Study Code	Control mean	Control std.dev	N (Control)	Treatment mean	Treatment std.dev	N (Treatment)
1	165	20	23	144	18	23
2	169	21	37	159	19	37
3	152	15	101	152	16	101
4	160	22	45	140	16	45
5	169	21	48	163	18	48
6	152	19	37	159	21	37
7	158	19	150	142	17	150
8	150	22	87	159	21	90

In order to enter your data you will first have to remove the data from the example given. Then hit “Calculate and Plot”. The underneath summaries come up.

<b>Summary Mean:</b>	<b>-6.672</b>
<b>Lower bound of 95% CI:</b>	<b>-8.886</b>
<b>Upper bound of 95% CI:</b>	<b>-4.458</b>
<b>Q statistics:</b>	<b>79.388</b>

Mean Difference (MD)	Pooled Variance	Weight	Product of Mean*Weight	sqdiffer	wt*sqdiff	MD_Hi	MD_Lo
-21.000	31.478	0.032	-0.667	205.288	6.522	-10.003	-31.997
-10.000	21.676	0.046	-0.461	11.075	0.511	-0.875	-19.125
0.000	4.762	0.210	0.000	44.517	9.348	4.277	-4.277
-20.000	16.444	0.061	-1.216	177.632	10.802	-12.052	-27.948
-6.000	15.938	0.063	-0.376	0.452	0.028	1.825	-13.825
7.000	21.676	0.046	0.323	186.927	8.624	16.125	-2.125
-16.000	4.333	0.231	-3.692	87.009	20.079	-11.920	-20.080
9.000	10.463	0.096	0.860	245.615	23.474	15.340	2.660

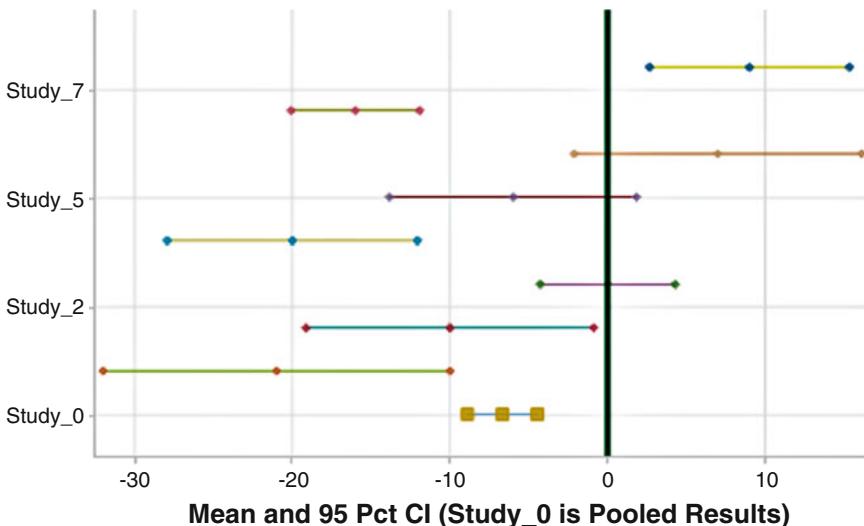
The chi-square value (Cochran Q statistic) is used by the online program for assessing fixed effect heterogeneity between the study outcomes. The chi-square

value equaled 79.388, and, thus, a p-value for (8-1) degrees of freedom of  $<0.001$  was obtained. Homogeneity had to be rejected.

The pooled result was very significant with a mean difference of  $-6.67$  (95% confidence interval ( $-4.46$  to  $-8.89$ , standard error =  $\pm 1.1$ ).

The t-statistic of  $-6.67/1.1 = 6\dots$ , with  $p < 0.001$ ). Active treatment performed much better than did control treatment. However, this result did not mean too much since the studies were too much heterogeneous for meaningful pooling.

Also, a graph of your meta-analysis can be drawn with the command: Plot Chart.



The mean differences per study with 95% confidence intervals are given. The pooled result shows, that the confidence interval is very small. The 95% confidence intervals of all of the studies are also given. They show a lot of heterogeneity as expected from the Q-statistic.

The above analysis is of course rather limited: and publication bias analysis and sensitivity analysis are missing. Only a fixed effect analysis of heterogeneity has been performed. Yet, the analysis is convenient, as, in a few seconds, your chief meta-data are readily produced from the internet. More sophisticated analyses will be in the next 20 or so chapters of the current edition.

We should add, that meta-analysis can be pretty easily performed using a pocket calculator (Clinical data analysis on a pocket calculator, Chap. 32, Meta-analysis of continuous data, Chap. 57, Meta-analysis of binary data, Springer Heidelberg Germany, 2015, from the same authors). The above online calculator does not provide meta-analyses of binary data. For the analysis of binary data we will apply the free MetaXL software, that can be automatically added to the Excel program of your computer, after downloading from the internet at [www.epigear.com](http://www.epigear.com).

## 5.4 Binary Outcome Data, MetaXL Free Meta-analysis Software

The free MetaXL meta-analysis software was used for analyzing a meta-analysis of 9 studies of the chance of death and infarction in patients with collateral coronary arteries compared to that in patients without collaterals (Akin et al., Effects of collaterals on deaths and re-infarctions in patients with coronary artery disease a meta-analysis, Neth Heart J 2013; DOI 10.1007). The data summaries per study are summarized underneath.

Study name	ES	Lo 95% CI	Hi 95% CI
monteiro	0,45	0,15	1,4
nathou	0,32	0,09	1,08
meier	0,37	0,25	0,56
sorajja	0,77	0,3	1,94
regieli	1,03	0,42	2,54
desch	0,3	0,11	0,81
steg	0,73	0,51	1,04
ter	0,45	1,15	1,4
cam	0,8	0,4	1,7

The effect size (ES) effect per study is the odds ratio per study (= odds of infarct with collaterals/odds of infarct without collaterals). The larger the odds ratio, the greater the chance of death/infarction. We will use MetaXL for data analysis.

The MetaXL software is downloaded by following a series of very simple commands, and after pressing Finish, you can observe that in the menu bar (upper horizontal bar of your opening Excel screen) the text MetaXL has been added. Click MetaXL, then click in the pop-up Examples. Seventeen real data examples are given, and you can check all of them or first study their characteristics in the free MetaXL User Guide version 5.1 that can be printed from [www.epigear.com](http://www.epigear.com). The example DiureticPreEc is pretty similar to our data: odds ratios and 95% confidence intervals are the main output. Now remove the data given and the names of the studies and enter subsequently the data of our collaterals meta-analysis. The result is given below.

Comparison of random effect and IVhet models, see sheet2 for forest plots

Cells with MetaXL functions are shaded red

Study name	ES	Lo 95% CI	Hi 95% CI
monteiro	0,45	0,15	1,4
nathou	0,32	0,09	1,08
meier	0,37	0,25	0,56
sorajja	0,77	0,3	1,94
regieli	1,03	0,42	2,54
desch	0,3	0,11	0,81
steg	0,73	0,51	1,04
ter	0,45	1,15	1,4
cam	0,8	0,4	1,7

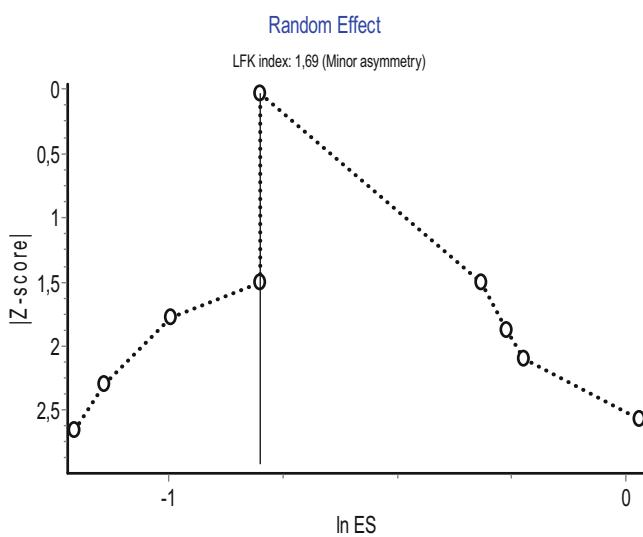
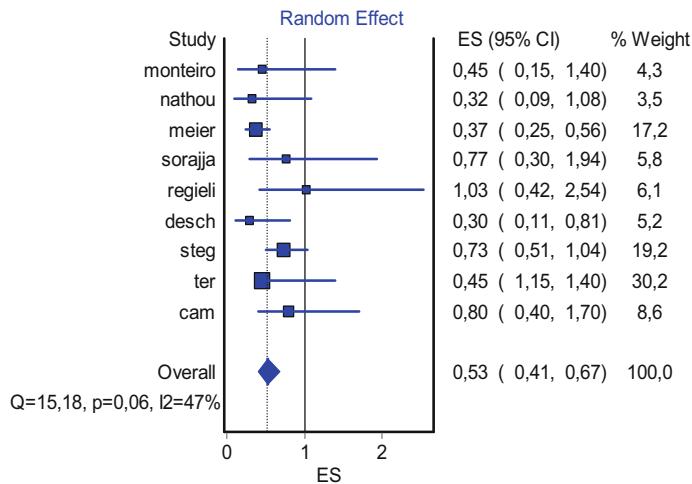
IVhet

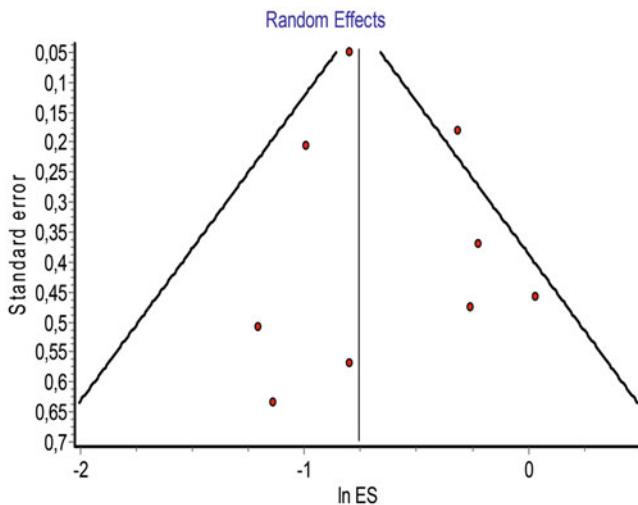
Random Effect

The odds ratios are very heterogeneous, and the fixed effect analysis should be similarly so. We will, therefore, as already commanded in the example, perform, respectively, a traditional random effect analysis and a quasi likelihood analysis for heterogeneity, entitled inversed variances heterogeneity (IVhet) analysis, which is a more robust alternative to the traditional Dersimonian random effect analysis.

#### 5.4.1 Traditional Random Effect Analysis

First the results of the random effect analysis will be listed. Click Results. The Choose an analysis window comes up. Click Random Effect. A Forest plot of the analysis is given.





The above forest plot (upper graph) shows, that, despite the heterogeneity in effect sizes, none of the studies produced an odds ratio  $>1$ . The 95% confidence intervals of the pooled odds ratio were expectedly very small (0.53, with 95% confidence intervals of 0.41–0.67).

This pooled odds ratio is significantly smaller than 1.00, because, after log transformation, the log odds ratio will be  $-0.63$  with 95% confidence intervals between  $-0.844$  and  $-0.400$ . The t -statistic as computed from the equation

$$\text{logodds ratio/its standard error} = -0.63 / -0.11 = 5, \dots$$

corresponds with a p value of  $<0.000$ .

The above Doi plot (middle graph) is expected to have a symmetric right and left limb with similar numbers of studies on either limb. With four of them on either side and one in the middle, the plot pretty much underscores, that publication bias is not a problem in this meta-analysis.

The above funnel plot, otherwise called Christmas tree plot, (lower graph) would have to show publication bias, if small studies with small effects had not been published. This phenomenon is, however, not obvious from the graph.

Random Effects results				
Options				
Forest plot	Doi plot	Funnel plot	Table	Exclude
Study	ES	LCI 95%	HCI 95%	weight (%)
monteiro	0,450	0,150	1,400	4,255
nathou	0,320	0,090	1,080	3,529
meier	0,370	0,250	0,560	17,236
sorajia	0,770	0,300	1,940	5,759
regieli	1,030	0,420	2,540	6,112
desch	0,300	0,110	0,810	5,152
steg	0,730	0,510	1,040	19,157
ter	0,450	1,150	1,400	30,244
cam	0,800	0,400	1,700	8,555
Pooled	0,526	0,411	0,674	100,000
Statistics				
I-squared	47,291	0,000	75,520	
Cochran's Q	15,178			
Chi2, p	0,056			
tau2	0,050			

The above table shows, that summary statistics will be separately given, if you click the term "Table" in the Options bar. "I-squared" =  $I^2$ , can be interpreted, as the proportion of total variation in meta-analysis due to heterogeneity, rather than sampling error. Fifty % is, generally, used as a cut-off for heterogeneity.

$$I^2 = 100\% * [Q - (k - 1)]/Q,$$

with  $Q$  = Cochran's  $Q$  and  $k$  = number of studies in a meta-analysis.

$I^2$  should be larger than 50% in order to conclude the presence of random heterogeneity. A Cochran's  $Q$  of over 15 with  $k-1 = 9-1$  degrees of freedom indicates, that homogeneity between the studies can not be definitely rejected because of a  $p$ -value of 0.056, and thus  $>0.050$ . The above  $\tau^2$  value is the estimated standard deviation of the differences between the studies due to a possible random effect. Considering the ranges of confidence intervals of the studies

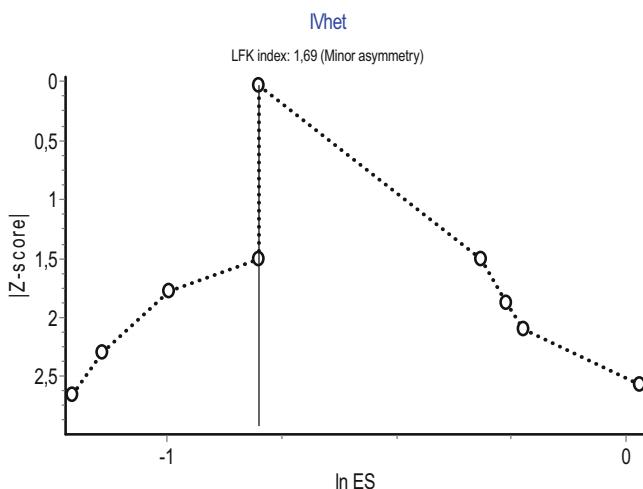
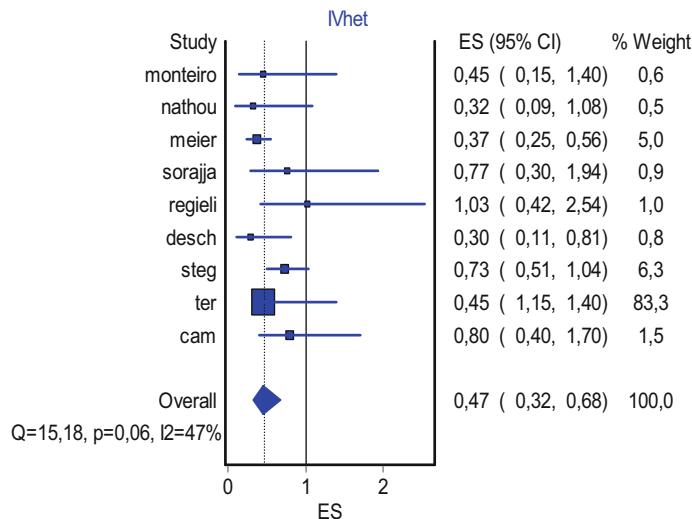
between 0.09 and 2.54, this is pretty small, but, unfortunately, a confidence interval is not given in the table.

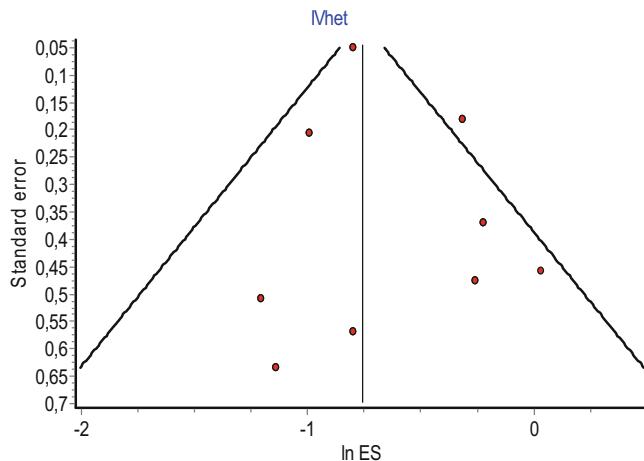
Random Effects results								
Options								
	Forest plot	Doi plot	Funnel plot	Table	Exclude	Sensitivity		
Excluded study	Pooled ES	LCI 95%	HC 95%	Cochran Q	Chi2	I <sup>2</sup>	I <sup>2</sup> LCI 95%	I <sup>2</sup> HC 95%
monteiro	0,533	0,408	0,695	15,173	0,034	53,886	0,000	79,202
nathou	0,538	0,415	0,699	14,819	0,038	52,762	0,000	78,765
meier	0,571	0,425	0,766	13,820	0,054	49,347	0,000	77,401
sorajja	0,515	0,398	0,666	14,069	0,050	50,245	0,000	77,762
rejeli	0,502	0,396	0,636	12,187	0,095	42,560	0,000	74,605
desch	0,544	0,420	0,706	14,413	0,044	51,434	0,000	78,237
steg	0,472	0,382	0,582	8,756	0,271	20,057	0,000	62,518
ter	0,563	0,406	0,781	11,750	0,109	40,425	0,000	73,690
cam	0,505	0,393	0,651	13,025	0,071	46,259	0,000	76,147

The above table will come up, if you click Exclude Sensitivity in the Options bar. The Chi<sup>2</sup> (chi square) column gives the p-values of the Cochran Q test statistics, after excluding one study at the time. Mostly results were pretty much unchanged, and the meta-analysis did, thus, not lack sensitivity, and was pretty much robust against the effects of one or more studies being of lower quality than the others.

#### 5.4.2 Quasi Likelihood (Invert Variance Heterogeneity (IVhet)) Modeling for Heterogeneity

The underneath graphs and tables are produced by the program if you command in the Options pop up window IVhet instead of random effect analysis. It is obvious that the results are virtually identical to those of the random effect analysis.





The IVhet forest plot is in the above upper graph. The IVhet doi plot is in the middle, and the IVhet funnel plot is in the lower graph. The interpretation is similar to that of the random effect graphs. The IVhet results as computed are in the underneath table.

IVhet results					
Options					
	Forest plot	Doi plot	Funnel plot	Table	Exclude
Study				ES	LCI 95%
monteiro				0,450	0,150
nathou				0,320	0,090
meier				0,370	0,250
sorajja				0,770	0,300
regieli				1,030	0,420
desch				0,300	0,110
steg				0,730	0,510
ter				0,450	1,150
cam				0,800	0,400
Pooled				0,467	0,320
Statistics					0,682
I-squared				47,291	0,000
Cochran's Q				15,178	75,520
Chi2, p				0,056	

With IVhet analysis no  $\tau^2$  is calculated.

IVhet results								
Options								
	Forest plot	Doi plot	Funnel plot	Table	Exclude	Sensitivity		
Excluded study	Pooled ES	LCI 95%	HCI 95%	Cochran Q	Chi 2	I <sup>2</sup>	I <sup>2</sup> LCI 95%	I <sup>2</sup> HCI 95%
monteiro	0,468	0,310	0,706	15,173	0,034	53,886	0,000	79,202
nathou	0,468	0,313	0,700	14,819	0,038	52,762	0,000	78,765
meier	0,473	0,301	0,744	13,820	0,054	49,347	0,000	77,401
sorajja	0,465	0,315	0,686	14,069	0,050	50,245	0,000	77,762
regieli	0,464	0,331	0,650	12,187	0,095	42,560	0,000	74,605
desch	0,469	0,316	0,697	14,413	0,044	51,434	0,000	78,237
steg	0,454	0,350	0,587	8,756	0,271	20,057	0,000	62,518
ter	0,565	0,397	0,804	11,750	0,109	40,425	0,000	73,690
cam	0,464	0,320	0,671	13,025	0,071	46,259	0,000	76,147

The IVhet analysis of sensitivity (= robustness) of this meta-analysis is above. The pooled effect size data were different from those of the random effect analysis, but the test statistics Q and chisquare were similar.

The results of the IVhet analysis were, thus, hardly different from those of the random effect analysis. Only a bit smaller pooled effect sizes and bit wider confidence intervals were observed, but test statistics were not different. Obviously, no features of the data were specified, that could successfully improve the fit of the data of that of the random effect analysis. It neither provided a worse fit. Both random and IVhet analysis did, unlike the fixed effect analysis (not shown), not convincingly demonstrate heterogeneity with p-values >0.05 and all of the I<sup>2</sup> values (read I<sup>2</sup> values) around 50%.

## 5.5 Conclusion

The results of the IVhet analysis are hardly different from those of the traditional random effect analysis. Only a bit smaller pooled effect sizes and bit wider confidence intervals are observed, but tests statistics were not different. Obviously, no features of the data were specified, that could successfully improve the fit of the data as compared to that of the random effect analysis. It neither provided a worse

fit. Both traditional random and IVhet effect analysis did, unlike the fixed effect analysis (not shown), not convincingly demonstrate heterogeneity with p-values >0.05 and I<sup>2</sup> values (read I<sup>2</sup> values), around 50%.

The basis of quasi-likelihood testing for heterogeneity is not straightforward. How does it work. Some features of the data are automatically specified, e.g., data being continuous or not, variability changes with averages, dependent or independent data, medians or means, and the effects on them from the remainder of the data available, etc. From that exercise by the software program a better fit model than that of a traditional overall data model is obtained. For example, in your data a higher variance in the center of your data is observed by the software program or response distributions may be skewed, or response variability may be different in some intervals from the average response. The software program will often conclude, that the data are (pseudo-) binomial, rather than continuous. It, then, changes the standard deviation (SD) of your mean into the formula appropriate for binomial analysis, with standard deviations computed from binomials like mean x (1-mean) terms.

Calculations will be pretty tough, but, for a computer, this is no problem.

As an alternative to IVhet analysis for heterogeneity, the so-called quality effects estimator for meta-analyses of heterogeneous studies has been proposed by Doi and Barendregt et al., the same group that initiated the quasi-likelihood methodology (Contemp Clin Trials 2015; 45: 123–9). Like the IVhet method, it should lead to a decreased overall measure of spread of the pooled effect size in a meta-analysis as compared to that of the traditional random effect method. In addition, it should maintain the nominal level of probability coverage, i.e., the probability that a computed 95% confidence is indeed 95% of the time the true area of data coverage. How does it work. The study outcomes in a meta-analysis are assessed against a predefined list of safeguards against bias, like, e.g.:

Was the target population defined or not?

Were the diagnostic criteria defined or not?

Were the study samples truly random or rather convenience samples?

...

...

The scores are subsequently converted into quality ranks, and the summary of ranks is used as a distribution model of weights in addition to the traditional sample size weight. It has been demonstrated by Doi and Barendregt that this way of modeling data is more conservative than the fixed effect model and less conservative than the traditional random effect model. Few studies are real data studies, and most of them involve simulated data studies. Nonetheless, this is another promising approach to the meta-analysis of heterogeneous studies.

## Reference

SPSS statistical software and SPSS Modeler have been applied, as a help in the majority of this edition's chapters, e.g., the chaps. 13–18 and 24. SPSS statistical software not only provides meta-analysis software as a help, but also for entire analyses. However, SPSS for entire meta-analyses is pretty hard, because nothing is in the menu, and everything depends on your syntax capacities.

# **Chapter 6**

## **Meta-analyses of Randomized Controlled Trials**

### **Convenience Samples and Other Limitations**

**Abstract** Double-blind placebo-controlled trials control for many types of biases, including selection bias, confounding, placebo effects, time effects, carryover effects, interactions etc. Are they, therefore, necessarily perfect? In this chapter eight reasons are given why this is not so. It is, nonetheless, a very interesting and generally very rewarding activity as we shall see. Three recently published meta-analyses from the authors are used as examples. The first meta-analysis showed robustness against the differences between parallel-group and crossover designs. The second showed that multiple outcome variables were helpful in many ways to answer the overall scientific question. The third showed that large meta-analyses of randomized controlled trials need not necessarily be tested for pitfall assessment.

#### **6.1 Introduction**

Double blind randomized controlled trials are assumed to have virtually no flaws, and to have as compared to observational studies only advantages. Specific advantages are, that they control for any type of bias, like confounding, placebo effects, time effects, carryover effects, interactions etc. However, In spite of all of this, are they necessarily perfect?

First, if you ever were involved in clinical trials as an investigator, you know about the load of possible errors when collecting, transcribing and entering the data onto your computer, that you must continually keep in mind.

Second, as for the protocol, they may be biased by *extreme* and overtly strict inclusion criteria (Understanding clinical data analysis, Chap.10, pp. 181–214, Springer Heidelberg Germany, 2016, from the same authors), and the use of random data that are not longer randomly distributed anymore due to the inclusion of convenience samples (from selected hospitals), and patients with cut-off characteristics (like cut-off laboratory values) (Chap. 43, Statistics applied to clinical studies 5th edition, Springer Heidelberg Germany, 2012, from the same authors).

Third, as for the data analyses, inadequate data cleaning is a well-recognized feature (Clinical trials in jeopardy, JAMA 2001; 286: 302–4, from the same authors).

Fourth, as for the statistical analysis, type I and II errors of finding respectively an effect where there is none and of finding no effect where there is one, may of course also cause biased conclusions of studies.

Fifth, a meta-analysis of randomized controlled trials involves not only the flaws of the trials, but also specific flaws of meta-analyses in general as reviewed in the Chaps. 1 and 2.

Sixth, as for randomized controlled trials, they may be controlled for randomness, but a meta-analysis of them is uncontrolled for random effect, which can only partly be adjusted by a random effect analysis, which, generally, has little power, at all events less power than a fixed effect analysis (Chap. 4).

Seventh, a special point with randomized controlled trials are the strict requirements, because of their experimental character, and of the suspicion of governments and ethic committees regarding conflict of interests between scientific and financial goals. Conflict of interest disclosures were given in only 2 of 29 meta-analyses, industry-ties were given in none (Cochrane Collaboration). How well do meta-analyses disclose conflicts of interests. [Cochrane.org](http://Cochrane.org) 2012–01–13). In a recent meta-analysis of 11 randomized controlled trials involving 13,833 patients from the Beta-blocker in heart failure collaborative group the latter information was fortunately given (Br Med J 2016; 353: 1855–60). Unrestricted grants were provided by Glaxo and Menarini, the manufacturers of carvedilol and nebivolol, that was the treatment of respectively 4248 and 2128 of the patients included in the meta-analysis.

Eighth, a final point is the inclusion of both double-blind parallel-group and crossover trials. Particularly, the crossovers have the risk of specific biases like treatment-by-period interaction, time effect, power loss due to a negative correlation with repeated measures (Chaps. 35 and 36, Statistics applied to clinical studies 5th edition, Springer Heidelberg Germany, 2012, from the same authors).

And, so, we have to conclude that even double blind randomized controlled trials is an error ridden activity, and that it would, therefore, be silly to assume, that meta-analyses of them is the pinnacle of scientific knowledge. It is, nonetheless, a very interesting and generally very rewarding activity as we shall see. At least this is so, if you have a sound prior hypothesis, and apply the scientific method (Chap. 3).

In the current chapter three recently published meta-analyses, from the same authors as this edition, are used as examples. The first meta-analysis showed robustness against the differences between a parallel-group and crossover designs. The second showed, how to handle multiple outcomes. The third showed that large meta-analyses of confirmatory trials do not always need pitfall assessments.

## 6.2 Example 1: Single Outcomes

---

International Journal of Clinical Pharmacology and Therapeutics  
Vol.50-No.7 / 2012(478-82)

Potassium treatment for hypertension in patients with  
high salt intake meta-analysis

Eric van Bommel<sup>1</sup> and Ton Cleophas<sup>1,2</sup>

*1Department Medicine, Albert Schweitzer Hospital, Dordrecht, The Netherlands, and 2European College Pharmaceutical Medicine, Lyon, France*

---

Van Bommel and Cleophas published the above meta-analysis.

1. Scientific question:

Does potassium decrease blood pressure in patients with high sodium intake.

2. Hypothesis:

Potassium tablets of patients with high sodium intake do not reduce blood pressure better than does placebo.

3. Null-Hypothesis testing:

A multiple groups unpaired t-test of two independent populations of hypertensive patients from multiple parallel-group studies is performed.

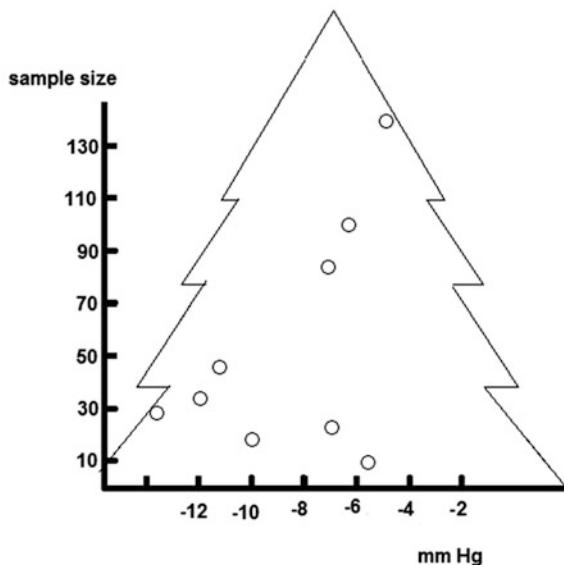
4. Result:

Potassium treatment better reduced high blood pressure in high sodium intake populations than did placebo.

	N	design	age	gender	salt intake	race
1.McGregor 1982 Lancet <sup>10</sup>	23	crossover	adult	m/f	high	20% blacks, 30% Asia (Charing Cross London)
2.Siani 1987 BMJ <sup>11</sup>	37	parallel	21-61	m/f	>170 mmol	Caucasians
3.Svetkey 1987 Hypertens <sup>12</sup>	101	parallel	51±12	m/f	high	blacks twice as many as the US average (Durham NC)
4.Krishna 1989 N Engl J Med <sup>13</sup>	10	crossover	adult	m	200 mmol	Caucasians
5.Obel 1989 J Cardiovasc Pharmacol <sup>1</sup>	48	parallel	20-60	m/f	>170 mmol	Blacks
6.Patki1990 BMJ <sup>14</sup>	37	crossover	49±8	m/f	192 mmol	Asians
7.Fotherby 1992 J Hypertens <sup>15</sup>	18	crossover	66-79	m/f	high	Caucasians with low renin hypertension consistent with high salt intake
8.Brancati 1996 Arch Int Med <sup>16</sup>	87	parallel	37-65	m/f	high	African Americans
9.Gu 2001 J Hypertens <sup>17</sup>	150	parallel	35-64	m/f	high	Chinese
10.Sarkkinen 2011 Nutr J <sup>18</sup>	45	parallel	25-75	m/f	high	Caucasians

Co-med = concomitant medication; m/f = male / female

The above table shows the characteristics of the 10 studies included after the exclusion of 32 studies that did not meet the above criteria for inclusion. Column 7 shows salt intake as estimated by the authors.



The above figure shows a Christmas tree plot of the systolic blood pressure reductions of the separate studies. Small studies with small results are obviously missing. This is compatible with the presence of some publication bias. No quantitative analyses were performed here, because of the small number of studies.

	N	difference systolic error	standard
1.McGregor 1982	23	-7.0	3.1
2.Siani 1987	37	-14.0	4.0
3.Svetkey 1987	101	-6.4	1.9
4.Krishna 1989	10	-5.5	3.8
5.Obel 1989	48	-41.0	2.6
6.Patki 1990	37	-12.1	2.6
7.Fotherby 1992	18	-10.0	3.8
8.Brancale 1996	87	-6.9	1.2
9.Gu 2001	150	-5.0	1.4
10.Sarkkinen 2011	45	-11.3	4.8

Pooled difference = -9.48 (95 % confidence interval -10.82 to -8.13)

Chi-square value = 206.9

p-value = < 0.0001

Heterogeneity chi-square value = 152.6, 9 degrees of freedom, p < 0.0001.

I-square value = 94.1% (< 50% cut-off for no heterogeneity).

The above table shows a pooled reduction of systolic blood pressure of  $-9.5 \text{ mm Hg}$  (95% confidence interval  $-10.8 \text{ to } -8.1$ ). This result was very heterogeneous with I-square values of 94%. Crossover studies are more at risk of biases than parallel-

group studies, e.g., time effects and carryover effects. Also single author may be more at risk of bias as multiple authors is commonly used as a quality criterion of clinical studies. Sensitivity assessments included separate outcome pooling of crossover and parallel-group, and outcome pooling after exclusion of the single authored studies. The results of the crossovers and the parallel-group studies were equal. However, after exclusion of single authored studies, the pooled result was less impressive, although it remained statistically statistically significant,  $-7.1 \text{ mm Hg}$  ( $-8.5$  to  $-5.7$ ). In the pooled systolic blood pressure data the presence of heterogeneity could now be rejected with a Thompson's I-square value of 24.3%. An I-square less than 50% means no heterogeneity in the meta-analysis (I-square is explained in the Chap. 2).

	N	difference systolic error	standard
1.McGregor 1982	23	-7.0	3.1
2.Siani 1987	37	-14.0	4.0
3.Svetkey 1987	101	-6.4	1.9
4.Krishna 1989	10	-5.5	3.8
6.Patki 1990	37	-12.1	2.6
7.Fotherby 1992	18	-10.0	3.8
8.Branca 1996	87	-6.9	1.2
9.Gu 2001	150	-5.0	1.4
10.Sarkkinen 2011	45	-11.3	4.8

Pooled difference =  $-7.12$  (95 % confidence interval  $-8.51$  to  $-5.72$ )

Chi-square value = 100.2

p-value =  $< 0.0001$

Heterogeneity chi-square value = 10.57, 8 degrees of freedom,  $p < 0.0001$ .

I-square value = 24.3% (< 50% cut-off for no heterogeneity).

The table shows a pooled reduction of systolic blood pressure of  $-9.5 \text{ mm Hg}$  (95% confidence interval  $-10.8$  to  $-8.1$ ). This result was very heterogeneous with I-square values of 94%. Crossover studies are more at risk of biases than parallel-group studies, e.g., time effects and carryover effects. Also single author may be more at risk of bias as multiple authors is commonly used as a quality criterion of clinical studies. Sensitivity assessments included separate outcome pooling of crossover and parallel-group, and outcome pooling after exclusion of the single authored studies. The results of the crossovers and the parallel-group studies were equal. However, after exclusion of single authored studies, the pooled result was less impressive, although it remained statistically statistically significant,  $-7.1 \text{ mm Hg}$  ( $-8.5$  to  $-5.7$ ). In the pooled systolic blood pressure data the presence of heterogeneity could now be rejected with a Thompson's I-square value of 24.3%. An I-square less than 50% means no heterogeneity in the meta-analysis (I-square is explained in the Chap. 2). In conclusion, in placebo-controlled clinical trials of

patients with high salt intake potassium supplementation significantly reduces the systolic blood pressure.

### 6.3 Example 1, Confirming the Scientific Question

The study confirmed the authors' prior scientific question and the null-hypothesis could be rejected. This is as expected, because it was based on sound clinical arguments. Sound clinical arguments included:

- (1) finding a better treatment for hypertension is vital, because hypertension is the first killer worldwide,
- (2) under-treatment is common,
- (3) hypertensive patients lack compliance with non drug treatments including salt intake reduction,
- (4) high salt intake causes hypertension.
- (5) potassium may be beneficial through the sodium-potassium cell pumps,
- (6) beneficial effects of potassium increasing drugs like potassium increasing diuretics, renin angiotensin blockers, and aldosterone inhibitors may be partly due to the same mechanism as potassium tablets,
- (7) potassium had cardiovascular benefits including the prevention of stroke, renal failure, and cardiac arrhythmias.
- (8) the DASH (Dietary Approaches to Stop Hypertension) trial prospectively tested in hypertensive patients a potassium – enriched diet.

### 6.4 Example 2: Multiple Outcomes

---

Ton J. Cleophas, Rob van Marum

p487–490

Published in issue: February 15 2001

[Abstract](#)[Full-Text](#) [HTML](#) [PDF](#)

EFFICACY AND SAFETY OF SECOND GENERATION DIHYDROPYRIDINE CALCIUM CHANNEL BLOCKERS IN HEART FAILURE, META-ANALYSIS  
T. J. CLEOPHAS, R. VAN MARUM, AMERICAN JOURNAL OF CARDIOLOGY

---

Cleophas et Van Marum published the above meta-analysis.

1. The scientific question:  
Does calcium channel blockade improve cardiac failure.
2. The scientific hypothesis:  
Calcium channel blockade does not improve cardiac failure.

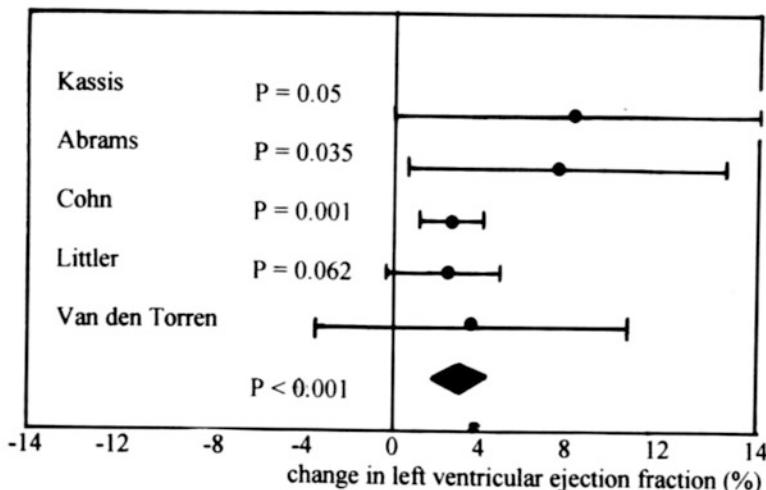
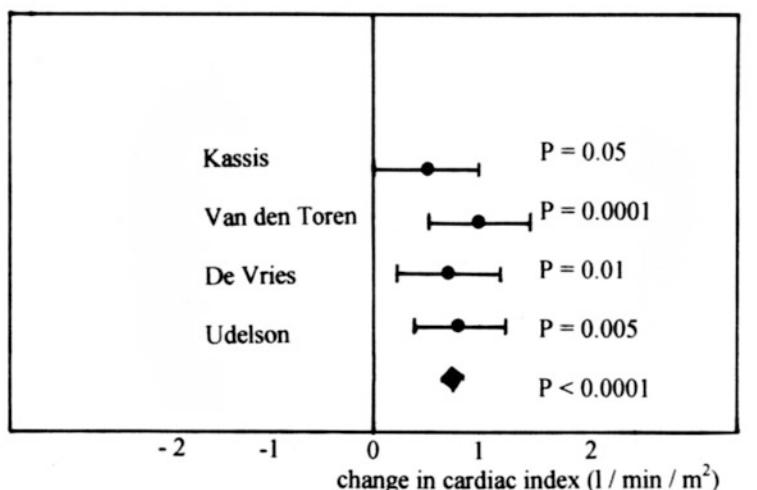
3. Null-hypothesis testing:

A multiple groups unpaired t-test of patients with cardiac failure on calcium channel blocker or placebo from multiple parallel-group studies is performed.

4. Result:

Calcium channel blockade better improved cardiac index, ejection fraction, and exercise tolerance than did placebo.

The study included 18 parallel-group studies of 3128 patients with CHF treated with second generation dihydropyridines or placebo. The effects of treatment on cardiac index, left ventricular ejection fraction, exercise treadmill duration, plasma norepinephrine level, and mortality were assessed.



- (1) The pooled cardiac index increased by 0.75 l/min/m<sup>2</sup> (95% CIs  $\pm$  0.10,  $P < 0.0001$ , see graph).
- (2) The pooled left ventricular ejection fraction increased by 2.5% (95% CIs  $\pm$  1.0,  $P < 0.001$ , see graph).
- (3) The pooled exercise tolerance time increased by 43 s (95% CIs  $\pm$  42,  $P = 0.05$ , no graph).
- (4) The pooled plasma norepinephrine levels fell by 45 pg/ml (95% CIs  $\pm$  56, not significant (NS), no graph).
- (5) Mortality was estimated in only 2 of the 18 studies, the PRAISE and the V-HeFT III studies: pooled OR – 0.94 (95% CIs –0.79 to 1.12, NS, no graph).

Fixed effects tests for heterogeneity were not statistically significant. Bonferroni adjustment for multiple testing produced a rejection p-value of  $0.05 \times (4/(5(5-1))) = 0.01$ . This would mean that the outcomes 3–5 were not statistically significant anymore. But, the outcomes 1 and 2 remained statistically very significant.

**Conclusions:** Second generation dihydropyridine calcium channel blockers significantly increase cardiac index, left ventricular ejection fraction, exercise tolerance test, and do not increase plasma norepinephrine levels in patients with CHF. And so, these drugs seem to be safe and beneficial to this category of patients. A 6% reduction in mortality was found. Although not significantly different from 0%, it does indicate that these drugs do not increase mortality in patients with CHF.

Fixed effects tests for heterogeneity were not statistically significant.

Bonferroni adjustment for multiple testing produced a rejection p-value of  $0.05 \times (4/(5(5-1))) = 0.01$ .

**Conclusions** Second generation dihydropyridine calcium channel blockers significantly increase cardiac index, left ventricular ejection fraction, exercise tolerance test, and do not increase plasma norepinephrine levels in patients with CHF. And so, these drugs seem to be safe and beneficial to this category of patients. A 6% reduction in mortality was found. Although not significantly different from 0%, it does indicate that these drugs do not increase mortality in patients with CHF.

## 6.5 Example 2, Handling Multiple Outcomes

In the example 2 five rather than a single outcome was tested. The multiple outcomes can be considered as a construct to help and confirm a single overall scientific question, i.e., do calcium channel blockers benefit patients with coronary heart failure? The Bonferroni adjustment for multiple outcome testing was applied.

Other methods for the purpose are possible.

1. Other adjustment procedures could have been, instead of Bonferroni which is over-conservative, meaning that p-values are rapidly too small and testing

becomes meaningless (Statistics applied to clinical studies 5th edition, Chap.<sup>9</sup> Multiple statistical inferences, Springer Heidelberg Germany, 2012, from the same authors).

2. A different more philosophical approach to the problem of multiple outcome variables is to look for trends without judging one or two low P-values among otherwise high P-values as proof. This requires discipline and is particularly efficient when multiple measurements are performed for the purpose of answering one single question, e.g., the benefit to health of a new drug estimated in terms of effect on mortality in addition to a number of morbidity variables. There is nothing wrong with this practice. We should not make any formal correction for multiple comparisons of this kind. Instead, we should informally integrate all the data before reaching a conclusion.
3. A further alternative for analyzing two or more primary variables is to design a summary measure or composite variable. With such an approach endpoint and primary variables must, of course, be assessed in advance, and the algorithm to calculate the composite must also be specified a priori. Since in this case primary variables are reduced to one composite, there is no need to make adjustments to salvage the type-I error rate.
4. Multivariate analysis of variance can be used to test simultaneously the level of significance of one outcome adjusted for other outcomes. In the Chap. 19 of this edition examples are given.

## 6.6 Example 3, Large Meta-analyses Without Need for Pitfall Assessment

---

Large Meta-Analyses of Randomized Controlled Trials Need Not Necessarily Be Tested for Biases

*Statistic Sessions, Eudipharm Lyon, December 2000*

A.H.Zwinderman, T.J.Cleophas, European Interuniversity College of Pharmaceutical Medicine, c/o Dept Medicine, Albert Schweitzer Hospital, Dordrecht, Netherlands

---

If meta-analyses include many trials that are adequately valid, they won't be very susceptible to the traditional biases of publication bias, heterogeneity, and lack of robustness. This paper is (1) to review, whether published meta-analyses comply with the standards to test for the above biases, (2) to test that in large meta-analyses the influence of such biases may be small and negligible. The meta-analyses of clinical trials published in 1998 in 4 general and in 8 specialist journals were included in the meta-analysis. A meta-analysis of 43 randomized controlled trials on the efficacy of angiotensin II antagonists (AII-r) for hypertension was used as example to test for publication bias, heterogeneity, and lack of robustness.

Four general journals published 45 meta-analyses all of whom complied with standards. The specialist journals published few meta-analyses, most of whom did not comply.

	Publication bias			
	trials n > 100	trials n < 100	difference	p-values
Fall in systolic blood pressure (mm Hg)	10.2 ± 0.2	10.8 ± 0.2	-0.6 ± 0.3	< 0.05
Fall in diastolic Blood pressure ( mm Hg)	8.1 ± 0.1	8.5 ± 0.2	-0.4 ± 0.2	< 0.05

Data are given as means ± standard errors of the mean (SEMs).

Heterogeneity	
Monotherapy	
Fall in systolic	
Blood pressure (range)	6.1 to 17.2 mm Hg *
Fall in diastolic	
Blood pressure (range)	4.0 to 13.4 mm Hg
Duplicate therapy	
Fall in systolic	
Bloodp ressure (range)	11 to 21.5 mm Hg *
Fall in diastolic	
blood pressure (range)	9 to 15.5 mm Hg

\*multiple groups ANOVA P < 0.001.

	Lack of robustness	
	Mean results SD/SEM studies	Mean results no SD/SEM studies (lower quality studies)
Low-dose data		
Fall systolic pressure (mm Hg)	10.8 ± 0.3	11.2 ± 0.3*
Fall diastolic pressure (mm Hg)	8.5 ± 0.1	8.9 ± 0.1**
High-dose data		
Fall systolic pressure (mm Hg)	13.3 ± 0.3	13.7 ± 0.3*
Fall diastolic pressure (mm Hg)	8.9 ± 0.2	9.9 ± 0.2***
Duplicate therapy data		
Fallinsystolicpressure (mm Hg)	13.3 ± 0.3	13.9 ± 0.3*
Fall diastolic pressure (mm Hg)	9.9 ± 0.2	10.8 ± 0.2***

\*P < 0.05 pooled data from low quality- versus those of high-quality trials.

\*\*P < 0.01

\*\*\*P < 0.001

Of 43 randomized controlled trials of a meta-analysis on AII-r for hypertension 23 included more than 100 patients. In these trials the average blood pressure fell by 10.2/ 8.1 mm Hg, in the trials with less than 100 patients by 10.8/8.5 mm Hg ( $0.6 \pm 0.3/0.4 \pm 0.2$ , both  $P < 0.05$ ), indicating the presence of some publication bias. The ranges of mean fall in blood pressures between the trials varied between 6 and 17.2 for systolic, and 4 and 13.4 mm Hg for diastolic blood pressures indicating the presence of substantial heterogeneity between the trials (both  $P < 0.001$ , multiple groups analysis of variance). Some trials were included that gave no Standard Deviations or Standard Errors SEMs) in their results. These studies, defined as of a lower quality than the rest, consistently gave larger mean reductions of blood pressures than the remainder (differences from 0.4 to 0.9 mm Hg,  $P < 0.05$  when weighted average SEMs of same drug and dose was used instead of missing SEMs), indicating some reduction in robustness in the meta-analysis. Despite the presence of a significant publication bias, heterogeneity, and lack of robustness, the overall effects of these factors were never larger than 5–6% of the treatment effects as measured.

It was concluded. Meta-analyses of controlled clinical trials published in specialist journal do not routinely test the presence of publication bias, heterogeneity between trials, and lack of robustness. The meta-analyses scrutinized in this paper although conducted according to appropriate standards, did neither. However, because the numbers of trials included were large enough to leave adequate power in the data, the effects of these factors were small, and so, the meta-analysis's main conclusions were not affected by this lack of adjustments. Large meta-analyses may not always have to be adjusted for publication bias, heterogeneity or lack of robustness.

## 6.7 Conclusion

Eight flaws of the otherwise called virtually unflawed method of double-blind placebo controlled clinical trials, are given for the benefit of future investigators applying this precious technique:

- (1) possible errors when collecting, transcribing and entering the data onto your computer,
- (2) protocols may be biased by *extreme* and overtly strict inclusion criteria, convenience samples, cut-off characteristics and laboratory values,
- (3) inadequate data cleaning is well-recognized,
- (4) type I and II errors,
- (5) meta-analysis of randomized controlled trials involving not only the flaws of the trials, but also specific flaws of meta-analyses in general like the meta-analysis pitfalls,
- (6) a meta-analysis is an uncontrolled exercise,
- (7) conflicts of interests,

(8) specific flaws with crossover trials (e.g., in the example 1).

Three real data examples of meta-analyses from placebo-controlled trials are given.

The first shows, that crossovers and parallel-group studies produced similar results, and can be used together in a meta-analysis.

The second shows, that many solutions exist for the trials with multiple outcomes.

The third shows, that large meta-analyses of confirmative trials need not necessarily be tested for the traditional pitfalls of meta-analyses.

## Reference

More information of convenience samples are in Statistics applied to clinical studies 5th edition, Chap. 43, Clinical trials do not use random samples anymore, pp 479–86, Springer Heidelberg Germany, 2012, from the same authors.

# **Chapter 7**

## **Meta-analysis of Observational Plus Randomized Studies**

### **Combined Meta-analysis of Different Classes of Study Designs**

**Abstract** This chapter reviews a publication from our group. It is the first publication of a combined meta-analysis of different classes of study designs. The real data example of 39,505 patients as used, showed, that combining the data from 10 observational and 7 randomized controlled trials provided more power of the pooled outcome, and a larger body of data enabling to consider a secondary outcome, i.e., excess in mortality.

Furthermore, it provided no publication bias, no lack of robustness, and little and clinically unimportant heterogeneity.

#### **7.1 Introduction and Example**

The flaws of observational studies include, that they do not provide random data, and, that mostly patients are recruited in the order of their outpatient clinic visit. Specific disadvantages of observational case-control studies are

- recall bias,
- underestimation of risk factors (severest patients never visit the outpatient clinic), and
- the presence of multiple differences between the sick and the controls.

Also a flaw of case-control studies is, that odds ratios are used as a surrogate for risk ratios, although they continually underestimate the true risk ratios.

Specific flaws of observational cohort studies are

- that they are lengthy, and,
- therefore, at risk of time effects,
- and that many differences in comorbidity exist between patients and controls.

In contrast, randomized controlled trials have relatively few flaws (Chap. 6). Specific strengths are, that they control for any type of bias, like

- confounding,
- placebo effects,
- time effects,

- carryover effects
- interactions etc.

An important weakness of randomized controlled trials is, that they are harder to be performed due to

- many rules including standardized protocols,
- signed informed consents,
- approval of scientific committees, and,
- conflict of interests and human experimentations law issues.

They are not only costly, but also continually at risk of being disapproved by ethic committees. Shrier et al., Am J Epidemiol 2007; 166: 1203–9, summarized many arguments, and came to conclude, that not only results were often similar, but also that meta-analyses including both observational and randomized studies outweighed the disadvantages as summarized above in many situations. To assess, whether meta-analyses of interventional studies should include observational, the underneath recently published example from our group was used.

J Card Fail **May 2009** Volume 15, Issue 4, Pages 305–309

Heart Failure Due to Systolic Dysfunction and Mortality in Diabetes: Pooled Analysis of 39,505 Subjects

**Masoor Kamalesh, MD**  **Ton J. Cleophas, MD, PhD**  
DOI: <http://dx.doi.org/10.1016/j.cardfail.2008.11.006>

## 7.2 Sound Clinical Arguments and Scientific Question

Meta-analyzing heart failure patients with versus without diabetes for mortality risk is relevant, but a randomized controlled clinical trial (RCT) protocol requiring to stop ACE inhibitors and beta-blockers for reason of symmetry in such patients is unethical.

As for observational research, another point is, that historically important scientific achievements have been obtained. A few examples of observational studies are given. They were not only accepted, but so much so, that they changed the world.

- *The case-control studies*
- “Cigarette smoking and bronchial carcinoma” (JAMA 1950), and
- “Contraceptive pill and heart infarct” (NEJM 1981),
- *the cohort study*
- “Lung cancer and smoking” (BMJ 1964).

Information based on observational studies may, therefore, improve information based on randomized trials. Both classes have their own strengths and weaknesses.

Observational studies report the real world, although they are at risk of confounding and placebo effects. Randomized trial protocols may adjust for such effects, but may never be approved, because of ethical and financial issues. An interesting approach is, to assume that one approach is complementary to the other, and that one approach may improve inferences based on the other. Kirtane et al. (Circulation 2009; [doi.org/10.1161](https://doi.org/10.1161)) studied the performance of drug eluting stents as compared to that of bare metal stents in both randomized and observational studies, and showed, that both methods were excellent, and, that the first one was significantly better only in one study type. However, a combined analysis was not performed in this meta-analysis. Benson and Hartz (N Engl J Med 2000; 342: 1878–86), similarly, compared observational and randomized controlled trials, but did neither combine the outcomes of the two methodologies. As the results of the observational studies produced estimates similar to those of the randomized controlled trials, a combined assessment seemed the next step.

In May 2009 our group published the first combined analysis of observational and randomized research, assessing the risk of death in patients with coronary heart failure in 39,505 patients with and without diabetes (J Card Fail 2009; 15: 305–9). This study will be used as example in the current chapter. In this meta-analysis the risk of death was smaller in the observational studies, than it was in the randomized controlled trials. This is unexpected, because, usually, results of studies are more spectacular the lower the quality. But the difference of the pooled risk ratios was small and not clinically relevant with a difference of <5%, and may, thus, be due to a type I error. There was no significant publication bias, and lack of robustness was not obvious. We came to conclude that combining randomized data and observational data is relevant, since

- consistent information of a large population may be obtained, and
- relevant secondary endpoints like morbidity statistics in a mortality meta-analysis underscoring the primary endpoint may be more easily obtained.

All of the other advantages as mentioned in the introduction, although maybe more theoretical and difficult to count, are, nonetheless, worthwhile to be considered here.

## 7.3 Summary Statistics

We meta-analyzed the data of 10 observational studies and 7 randomized trials, studying the death risk in heart failure with versus without diabetes. The patient data of the separate studies are in the underneath table. Studies were very different in magnitude.

**Table 1** Studies Reporting Heart Failure Mortality in Diabetes Included in Analysis

Author	Year	n	No. with diabetes	Follow-up (months)	Excess Mortality (%)	Comments
Dries et al <sup>7</sup>	2001	6797	647	37	37	Excess risk only in ischemic etiology <sup>7</sup>
Bobbio et al <sup>8</sup>	2003	3091	621	12	44	β-blocker benefit lower in diabetics <sup>†</sup>
Domanski et al <sup>9</sup>	2003	2708	975	24	33	Increased risk in ischemic group only
Gustafsson et al <sup>10</sup>	2004	5491	900	78	50	Risk much higher for women than men <sup>‡</sup>
De Groote et al <sup>11</sup>	2004	1246	274	40	54	Increased risk in ischemic group only
Murcia et al <sup>12</sup>	2004	2231	496	42	39	All patients were post-myocardial infarction <sup>§</sup>
Garcia et al <sup>20</sup>	2004	362	143	12	62	High prevalence of DM (40%) in CHF
Smooke et al <sup>13</sup>	2005	554	132	24	65	Risk higher (450%) if insulin treated
Gorelik et al <sup>14</sup>	2005	385	176	60		Risk even higher for women
Burger et al <sup>15</sup>	2005	498	236	6	78	Acute heart failure patients
V-Roman et al <sup>16</sup>	2005	1659	431	144	43	Increased risk only in nonischemic group <sup>¶</sup>
Deedwania et al <sup>17</sup>	2005	3991	985	12	8	Risk higher (26%) in severe heart failure <sup>§</sup>
Kamalesh et al <sup>18</sup>	2006	495	293	32	73	Predominantly male patients
From et al <sup>19</sup>	2006	665	133	60	48	Increased risk only in nonischemic group
Ahmed <sup>21</sup>	2007	7788	2218	38	31	Increased risk in women
Ghali <sup>22</sup>	2007	1520	622	16	-7	Cardiac resynchronization therapy used
Tribouilloy <sup>23</sup>	2008	386	96	60	66	Preserved systolic function—all subjects

## 7.4 Pooled Results

The underneath tables show pooled odds ratios of the randomized data, the risk ratios used as surrogates for the odds ratios of the observational studies, respectively 1.35 and 1.22.

**Table 2** Death risk of Diabetics Versus Nondiabetics; Meta-analysis of 7 Clinical Trials

<b>Study</b>	<b>Study Size</b>	<b>Odds Ratio</b>	<b>95% Confidence Interval</b>	<b>P Value</b>
1. Dries	6797	1.28	1.10–1.50	<.002
2. Bobbio	2834	1.44	1.16–1.78	<.05
3. Gustafsson	5491	1.50	1.30–1.70	<.001
4. Murcia	2231	1.39	1.14–1.68	<.01
5. Deedwania	3991	1.08	0.80–1.47	NS
6. Ahmed	4112	1.28	1.19–1.38	<.001
7. Ghali	1520	0.93	0.59–1.47	NS
Pooled		1.35	1.27–1.43	<.0001
Test heterogeneity NS*				

**Table 3** Death Risk of Diabetics Versus Nondiabetics; Meta-analysis of 10 Observational Studies

<b>Study</b>	<b>Study Size</b>	<b>Risk Ratio</b>	<b>95% Confidence Interval</b>	<b>P Value</b>
1. Garcia	362	2.90	1.30–6.30	<.01
2. Domanski	2708	1.22	1.06–1.41	<.02
3. De Groote	1246	1.06	0.80–1.41	NS
4. Smoode	554	2.64	1.39–5.00	<.02
5. Burger	498	1.78	1.19–2.65	<.01
6. Gorelik	385	1.09	1.00–1.19	<.05
7. Varela	1659	1.43	1.13–1.80	<.02
8. Kamalesh	495	1.70	1.16–2.51	<.01
9. From	665	1.33	1.07–1.66	<.02
10. Tribouilloy	386	1.80	1.27–2.48	<.01
Pooled		1.22	1.14–1.30	<.0001
Test heterogeneity $P < .001$ *				

The pooled overall risk ratio of all of the 17 studies was 1.28, and significantly heterogeneous as shown in the underneath table. This heterogeneity was explained by the observational studies with risk ratios of 1.06–2.90, while those of the randomized studies were 0.93–1.50. And the heterogeneity disappeared after removal of the observational studies. It was probably mainly caused by two (small) outlier studies: Garcia ( $n = 362$ , risk ratio 2.90) and Smooke ( $n = 554$ , risk ratio 2.64). The risk ratio of the combined analysis given underneath was 1.28, and was, thus, only slightly different that of the observational studies, 1.22, difference  $< 5\%$  (4.9%).

**Table 4** Death Risk of Diabetics Versus Nondiabetics; Combined Analysis of 7 Clinical Trials and 10 Observational Studies

	Risk Ratio	95% CI	P Value
Pooled	1.28	1.23–1.34	<.0001

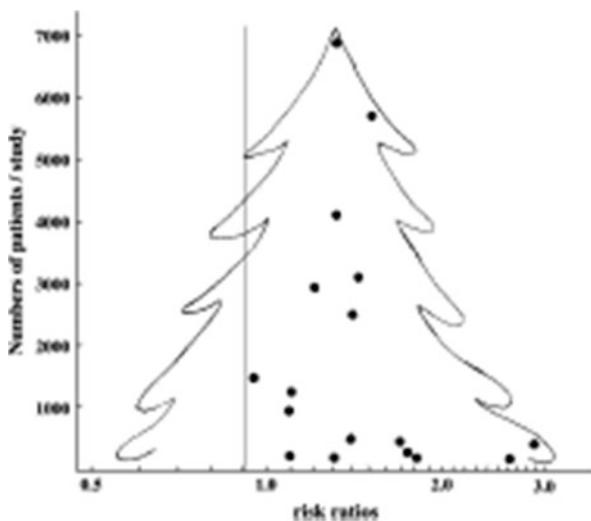
Test for heterogeneity  $P < .001$ .

## 7.5 Heterogeneity Assessments

Although statistically heterogeneous, clinically this heterogeneity was considered to be small and unimportant, and it was concluded, that the aggregate data confirmed the increased risk of mortality in patients with heart failure and diabetes as suggested, by some previous reports, and that it did so by 1.28 (95% confidence interval 1.23–1.34). The validity of this result was supported by negative tests both for publication bias and for lack of robustness.

## 7.6 Publication Bias Assessments

The figure below shows a Christmas tree plot suggesting the presence of some publication bias: small studies with negative results were presumably not published. However, the differences between the pooled risk ratio of the 2 large and the 13 small studies were not significantly different from one another: RR 1.40 and 1.23 (95% CI 1.27–1.54 and 1.17–1.32, respectively,  $t$  value of difference 1.83,  $P = .10$ ). A significant publication bias was thus not demonstrated.



## 7.7 Robustness Assessments

To assess robustness, the death RR of the clinical trials was compared with that of the observational studies under the assumption, that the observational studies were scientifically of less quality, than the clinical trials, and, therefore, would tend to produce results at higher levels of significance. The magnitude of the increased death risk, however, was not larger in the latter compared to that of the former studies (RR 1.38 and 1.22; 95% CI 1.27–1.49 and 1.14–1.30, respectively). Lack of robustness was, thus, not obvious.

## 7.8 Improved Information from the Combined Meta-analysis

The underneath table gives the pooled result of a secondary endpoint, the hospitalization risk, from the data of 4 randomized and 2 observational studies as available in the meta-data. The secondary endpoint underscored the validity of the conclusions from the primary endpoint, the risk of death.

**Table 5** Hospitalization Risk in Diabetics Versus Nondiabetics; Meta-Analysis of 4 Clinical Trials (1-5) and 2 Observational Studies (6,7)

<b>Study</b>	<b>Study Size</b>	<b>Risk Ratio</b>	<b>95% Confidence Interval</b>	<b>P Value</b>
1. Dries	6797	1.52	1.18–1.96	<.01
2. Bobbio	2843	1.28	1.11–1.49	<.002
3. Murcia	2231	1.65	1.35–2.01	<.05
4. Deedwania	3991	1.76	1.38–2.23	<.002
5. Ahmed	4112	1.28	1.19–1.38	<.001
6. Domanski	2708	1.16	1.02–1.32	<.001
7. Garcia	362	2.63	1.54–4.48	<.001
Pooled*		1.36	1.25–1.47	<.0001

## 7.9 Conclusion

The effects of addition to randomized studies of observational studies in a single meta-analysis was the subject of this chapter. The real data example of 39,505 patients used, showed that combining the data from 10 observational and 7 randomized controlled trials:

- (1) provided more power and more robust numbers,
- (2) provided a larger body of data enabling to consider possible reasons for excess in mortality,
- (3) provided no publication bias,
- (4) provided no lack of robustness,
- (5) provided little and clinically unimportant heterogeneity.

## Reference

More information of the meta-analysis of randomized controlled trials is in the Chap. 6, that of the meta-analysis of observational studies is in the Chap. 8.

# **Chapter 8**

## **Meta-analysis of Observational Studies**

### **Meta-analyzing Rare Diseases**

**Abstract** Controlled long term open evaluation studies will be applied, if clinical trials are not feasible. Particularly, for the study of rare events, such studies have been published.

In the current chapter two open evaluation meta-analyses from our group will be reviewed. The first meta-analysis was homogeneous and robust. The scientific method was helpful to confirm its prior scientific question. The second meta-analysis was different. It only included studies either performed by internists or by pharmacists. The null-hypothesis was: no difference in outcome between one study and the other. In this meta-analysis heterogeneity was a benefit rather than pitfall, because the null-hypothesis was no heterogeneity, that, hopefully, could be rejected.

### **8.1 Introduction**

Controlled long term open evaluation studies will be often applied, if clinical trials are not feasible. Particularly, for the study of rare events like adverse drug effects or in rare subgroups, like myocardial infarctions in patients with collateral coronary arteries, or iatrogenic admissions to hospitals, only such studies have been published. These studies are prospective, minimizing the risk of recall bias and other flaws of retrospective studies including risk factor underestimation, because the severest patients never visited the clinic, and the presence of multiple confounders, because of the many differences between the sick and the controls. Open evaluation studies, albeit explorative in nature, give rise to relevant conclusions. In addition, authors of any type of study design try and report the most unbiased version of their study, irrespective of its design. This explains, why meta-analyses of studies of any design are not impossible, as long as there are controls. It is, of course, comprehensible, that studies with lower quality are more at risk of pitfalls like lack of robustness and homogeneity. In this chapter two recently completed meta-analyses of open evaluation studies from our group are reviewed.

Despite its observational nature, the first meta-analysis was homogeneous and robust. The scientific method was helpful to confirm its prior scientific question. With strong predictors, particularly the traditional risk factors of coronary artery

disease, it has been advocated to always include them in the assessment of other predictors (Cleophas, Angiology 1996; 47: 789–96). Therefore meta-regressions were performed with the odds ratios of diabetes, hypertension, cholesterol, and smoking as predictors and the the odds ratios of collateral coronary arteries as outcome. The second meta-analysis was different. It consisted of studies performed either by internists or by pharmacists. The null-hypothesis was no difference in outcome between one study and the other. In this meta-analysis heterogeneity was a benefit rather than pitfall, because the null-hypothesis was no heterogeneity, that, hopefully, could be rejected.

## 8.2 Prospective Open Evaluation Studies

Prospective open evaluation studies will be often applied if clinical trials are not feasible. E.g., for the study of rare events like adverse drug effects or in rare subgroups, like myocardial infarctions in patients with collateral coronary arteries, only open evaluation studies have been published. This means that randomness, placebo-effects, and all of the flaws of observational studies like time effects, selection bias, carryover effects etc. have to be taken into account, in addition to the traditional pitfalls of meta-analyses, including publication bias, heterogeneity, lack of robustness. Fortunately, these studies are prospective, minimizing the risk of recall bias and other flaws of retrospective studies including risk factor underestimation, because the severest patients never visited the clinic, and the presence of multiple confounders, because of the many differences between the sick and the controls (see Chap. 2, Randomized and observational research, pp. 11–27, in: Understanding clinical data analysis, Springer Heidelberg Germany, 2016, from the same authors). Nonetheless, explorative analyses of observational research gives rise to relevant conclusions, that can be confirmed in future studies. In addition, authors of any type of study try and report the most unbiased version of their study, irrespective of its design, although, with the studies included in agenda-driven meta-analyses, this may not be entirely true (Stegenga, Stud Hist Philos Biol Biomed Sci 2011; 42: 497–507). This explains, why meta-analyses of studies of any design are possible, as long as there are controls. It is, of course, understandable, that studies with lower quality are more at risk of lack of robustness and homogeneity. This chapter is the second of three chapters covering meta-analyses of both controlled and observational clinical trials.

### 8.3 Example 1, Event Analysis in Patients with Collateral Coronary Arteries

---

2012; 21: 146-51 · Netherlands heart journal: monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation  
Effect of collaterals on deaths and re-infarctions in patients with coronary artery disease: A meta-analysis

S Akin · T Yetgin · J J Brugts · A Dirkali · F Zijlstra · T J Cleophas

---

A total of 9 studies describing deaths enrolling 6791 participants were included. Studies from the year 2000 until 2012 were included. Medline, Google, major journals, and PubMed were searched, and reference lists of selected articles on the subject.

All studies containing information on the presence of collaterals according to Rentrop's criteria or positive blush tests from the collaterals into the ischemic area were included. The endpoint of this analysis was the impact of collateral circulation on all-cause mortality and the combinations "all-cause deaths and re-infarctions". It was assumed that a loglinear relationship existed between the time of follow up and the odds of events, and that patient stratification with blush and Rentrop's criteria would produce similar patterns. The presence of traditional atherosclerotic risk factors in the patients with and without collaterals were assessed with odds ratios, and their effects on deaths and the composite "death and re-infarctions" were assessed with meta-regression using a multiple linear regression model with the odds ratios of the risk factors as predictors and the odds ratios of deaths and the composite "deaths and re-infarctions" as outcome.

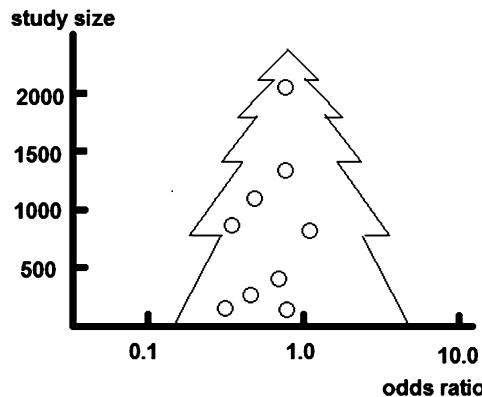
Fifteen studies were, initially, included. Five of them were excluded, because no survival data were reported per group. Of the ten studies included nine studies reported deaths during follow-up 0.6–9 years, while 7 studies reported the composite endpoint "deaths or recurrent infarction" separately for the subgroups.

### 8.4 Example 1, the Scientific Method

Meta-analysis of coronary events and collaterals.

1. Scientific question:  
Do coronary collaterals protect against coronary events.
2. Hypothesis:  
Coronary collaterals do not protect.
3. Null-Hypothesis testing:  
A multiple groups unpaired chi-square test tests the odds ratios of coronary events in patients with collaterals versus those without.
4. Result:  
Patients with coronary collaterals are less at risk of myocardial infarction than those without.

## 8.5 Example 1, Publication Bias



The above figure shows a Christmas tree plot: small studies with large odds ratios are not seen. This could mean they are at risk of not being published, and, thus, suggests the presence of some publication bias.

## 8.6 Example 1, Pooled Results, Tests for Heterogeneity and Robustness

	Odds Collaterals	odds no collaterals	odd ratio	t-value	p
1.Antoniucci	11/253	81/819	0.44	-2.50	0.0125
2.Monteiro	4/31	6/29	0.62	-0.68	0.497
3.Elsman	3/103	45/908	0.59	-0.88	0.380
4.Meier	25/201	170/416	0.30	-5.15	0.0001
5.Sorajja	3/116	8/191	0.62	-0.70	0.483
6.Regigli	2/261	4/612	1.17	+0.18	0.855
7.Desch	3/66	22/144	0.30	-1.92	0.056
8.Steg	155/1767	28/2232	0.70	-1.65	0.098
9.Ilia	3/44	7/27	0.26	-1.83	0.068
Pooled odds ratio			0.47	-5.90	0.0001

Heterogeneity chi-square value = 9.1724, 8 degrees of freedom, not significant.

$I^2$ -square value = 12.8 % (< 50 % cut-off for no heterogeneity).

The above table shows, that the pooled odds ratio of 9 studies equaled 0.47, meaning that patients with collaterals have a more than twice reduced risk of dying during a follow-up period of up to 9 years. The validity of this finding is supported by the lack of heterogeneity and an adequate  $I^2$  square value.

	Odds Collaterals	odds no collaterals	odd ratio	t-value	p
1.Antoniucci	11/253	81/819	0.44	-2.50	0.0125
2.Monteiro	4/31	6/29	0.62	-0.68	0.0497
3.Elsman	3/103	45/9081	0.59	-0.88	0.380
5.Sorajja	3/116	8/191	0.62	-0.70	0.483
6.Regieli	2/261	4/612	1.17	+0.18	0.855
7.Desch	3/66	22/144	0.30	-1.92	0.056
9.Ilia	3/44	7/27	0.26	-1.83	0.068
Pooled odds ratio			0.47	-3.50	0.0005
Heterogeneity chi-square value = 2.3064, 6 degrees of freedom, not significant.					
I <sup>2</sup> -square value = 0.00 % (< 50 % cut-off for no heterogeneity).					

The above table assesses robustness of the meta-analysis. After exclusion of two very asymmetric studies the pooled odds ratio is unchanged, indicating, that the current meta-analysis is robust against the bias of asymmetry in the given studies. After exclusion of the asymmetric studies the pooled confidence intervals were a little bit wider and the t-value and p-value was a bit larger and smaller respectively, indicating some loss of power due to a smaller sample size left in the meta-analysis, but, otherwise, the pooled results were unchanged.

	Odds Collaterals	odds no collaterals	odd ratio	t-value	p
1.Monteiro	6/29	11/24	0.45	-1.38	1.69
2.Nathou	3/173	20/365	0.32	-1.84	0.066
3.Meier	36/190	197/389	0.37	-4.87	0.0001
4.Sorajja	7/112	15/184	0.77	-0.56	0.576
5.Regieli	7/254	16/600	1.03	+0.07	0.944
6.Desch	5/64	34/132	0.30	-2.38	0.018
7.Steg	246/1676	42/2092	0.73	-1.72	0.085
Pooled odds ratio			0.54	-5.22	0.0001
Heterogeneity chi-square value = 10.7317, 6 degrees of freedom, 0.05<p<0.10.					
I <sup>2</sup> -square value = 44.0 % (< 50 % cut-off for no heterogeneity).					

The above table shows a meta-analysis with the composite “death and recurrent infarction” as endpoint. The odds ratio of “death and recurrent infarction” between those with collaterals versus those without was 0.54. Again heterogeneity according to the fixed effects test and the I-square value were small, although a trend to heterogeneity was observed. After the exclusion of the asymmetric studies the odds ratio somewhat rose to 0.60, and, at the same time, no more heterogeneity was observed. This suggests, that the asymmetric characteristics may have contributed to a lack of homogeneity, and that it was not entirely robust against this potential flaw. A significant reduction of the composite “deaths and recurrent infarction” with an odds ratio of 0.60 was observed in the analysis, albeit with only 4 studies left in the meta-analysis as shown underneath.

	Odds Collaterals	odds no collaterals	odd ratio	t-value	p
1.Monteiro	6/29	11/24	0.45	-1.38	0.169
4.Sorajja	7/112	15/184	0.77	-0.56	0.576
5.Regieci	7/254	16/600	1.03	+0.07	0.944
6.Desch	5/64	34/132	0.30	-2.38	0.017
Pooled odds ratio			0.60	-2.03	0.043
Heterogeneity chi-square value = 3.754, 3 degrees of freedom, ns.					
I <sup>2</sup> -square value = 20.1 % (< 50 % cut-off for no heterogeneity).					

## 8.7 Example 1, Meta-regression Analysis

With strong predictors, like the traditional risk factors of coronary artery disease, it has been advocated to always include them in assessments of other predictors (Cleophas, Angiology 1996; 47: 789–96). The underneath table shows the odds of the presence of traditional risk factors in patients with collaterals versus those without collaterals in the various studies. Linear meta-regressions with the odds ratios of the traditional risk factors as independent and those of deaths and the composite endpoint “deaths plus re-infarctions” as dependent variable were performed.

Odds ratios in patients with collaterals versus those without

	Diabetes	hypertension	cholesterol	smoking
1.Monteiro	1.61	1.12	2.56	0.93
2.Elsman	0.62	1.10	1.35	0.93
3.Meier	1.13	0.69	1.33	1.85
4.Nahoe	0.76	0.85	1.34	0.78
5.Sorajja	1.69	0.83	1.11	1.09
6.Regieł <sup>*</sup>	1.02		1.28	
7.Steg	0.13	0.17	0.21	0.27
8.Desch	1.52	0.79	0.85	1.25
9.Ilia	0.65	0.74	1.04	0.83

Meta-regression between odds ratios of risk factors and those of deaths.

F-value	0.134	0.242	0.022	0.204
P-value	0.733	0.640	0.889	0.667

Multiple regression with all of the risk factors tested simultaneously

F-value	0.442
P-value	0.780

Meta-regression between odds ratios of risk factors and those of the composite endpoint deaths plus re-infarctions.

F-value	0.134	0.007	0.587	0.156
P-value	0.733	0.938	0.486	0.709

Multiple regression with all of the risk factors tested simultaneously

F-value	0.523
P-value	0.761

---

<sup>\*</sup>No qualitative data of diabetes and cholesterol were reported in this study.

The calculated odds ratios is given of the presence of traditional atherosclerotic risk factors in the patients with and without collaterals, and their linear relationships with the odds ratios of death and “deaths plus re-infarct”. No significant relationships were observed. When the composite “deaths and re-infarct” was used similarly no significant relationships were observed. This would mean that the increased risk of death with collaterals is not caused by an increased risk of the presence of the traditional risk factors in the subgroups of patients with collaterals.

## 8.8 Conclusions

1. In 6791 CAD patients from the post-PCI era the presence of collaterals reduced deaths by 0.47 ( $p < 0.0001$ ) and “death plus re-infarctions by 0.60 ( $p = 0.043$ ).
2. Many studies in the past were negative due confounding as a consequence of asymmetric patient characteristics.
3. In the present meta-data the atherosclerotic risk factors were no more present in the patients with collaterals than they were in those without.

## 8.9 Example 2, Event Analysis of Iatrogenic Hospital Admissions

---

International Journal of Clinical Pharmacology and Therapeutics  
Vol.47-2009(549-56)

**Prevalence of iatrogenic admissions to the departments of medicine / cardiology / pulmonology in a 1250 beds general hospital**  
Roya Atiqi, Erik van Bommel, Ton J. Cleophas, Aeilko H. Zwinderman

Earlier studies on patients admitted for adverse drugs effects (ADEs) were very heterogeneous: percentages varied from 1.0–16.8% (Lazarou et al., JAMA 1998; 279: 1200–5). In the present paper the percentages of patients admitted to hospital due to ADEs in recent years were assessed. In this meta-analysis, rather than homogeneity, the null-hypothesis was heterogeneity. The investigators were hopeful to reject homogeneity.

## 8.10 Example 2, the Scientific Method

Meta-analysis of adverse drug effect admissions and the type of research group is given.

1. Scientific question:  
Is the type of research group a determinant of the numbers of adverse drug effect admissions to hospital.
2. Hypothesis:  
The type of research group is not a determinant.
3. Null-Hypothesis testing:

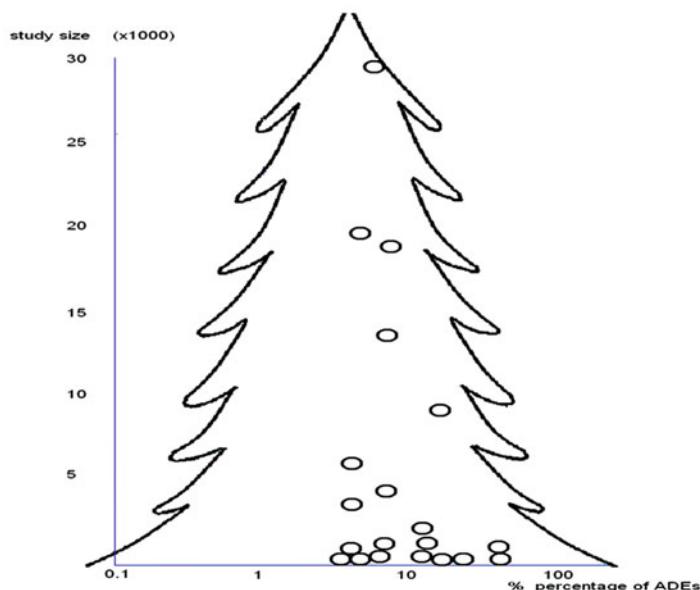
In a linear regression controlled for age and study size it is assessed whether pharmacists report adverse drug effects more often than internists do.

#### 4. Result:

The difference was very significantly so, the null hypothesis was rejected.

## 8.11 Example 2, Publication Bias

Publication bias was assessed by plotting on a semi-logarithmic scale the percentages of ADEs of the individual studies against the sample sizes of the individual studies.



The above figure suggests that larger studies tend to have less spectacular results, and that there seems to be a Christmas-tree-like pattern (funnel-pattern). The smaller the studies, the larger the percentage of adverse drug effects. The graph is also suggestive of publication bias, because small studies with small results are largely missing, and may thus not have been published.

## 8.12 Example 2, Overall Results and Heterogeneity and Lack of Robustness

Study	study size	percentage of all admissions	95% confidence intervals
1.Mannesse et al 2000	106	21.0 %	13.0-29.0 %
2. Malhotra et al 2001	578	14.4 %	11.5-17.3 %
3. Chan et al 2001	240	30.4 %	24.5-36.3 %
4. Olivier et al 2002	671	6.1 %	4.3-7.9 %
5. Mjordal et al 2002	681	12.0 %	9.5-14.5 %
6. Onder et al. 2002	28411	3.4 %	3.2-3.6 %
7. Koh et al 2003	347	6.6 %	4.0-9.2 %
8. Easton-Carter et al 2003	8601	3.3 %	2.9-3.7 %
9. Dormann et al 2003	915	4.9 %	4.9-14.3 %
10.Peyriere et al 2003	156	9.6 %	4.9-14.3 %
11.Howard et al	4093	6.5 %	5.7-7.3 %
12. Pirmohamed et al	18820	6.5 %	6.2-6.8 %
13.Hardmeier et al 2004	6383	4.1 %	3.6-4.6 %
14.Easton et al 2004	2933	4.3 %	3.6-5.0 %
15.Capuano et al. 2004	480	3.5 %	1.9-5.1 %
16.Caamano et al 2005	19070	4.3 %	3.7-4.6 %
17.Yee et al 2005	2169	12.6 %	11.2-14.0 %
18.Baena et al. 2006	2261	33.2 %	31.2-35.2 %
19.Van den Bemt et al 2006	12793	5.6 %	5.2-6.0 %
20.Van der Hooft et al 2008	355	5.1 %	2.8-7.4 %
Pooled	113203	5.2 %	5.0-5.4 %

The above table shows the individual studies included in the meta-analysis. The pooled result of the 20 studies as included provided an overall percentage of ADEs of 5.2% (95% confidence interval 5.0–5.4%). This pooled result was not significantly different from that of 21 studies published before 1997, as shown underneath.

Source	study size	percentage of all admissions	95% confidence interval
1. Smith et al 1966	900	1.7 %	0.9-2.5 %
2. Seidle et al 1966	714	3.9 %	2.5-5.3 %
3. Sidel et al 1967	267	4.5 %	2.1-6.9 %
4. Gardner and Watson 1970	939	5.1 %	3.7-6.5 %
5. McKenzie et al 1973	658	2.9 %	1.6-4.5 %
6. Miller 1974	2065	2.9 %	2.2-3.6 %
7. Miller 1974	1193	5.6 %	4.3-6.9 %
8. Miller 1974	1025	3.0 %	1.9-4.1 %
9. Miller 1974	555	1.8 %	0.7-2.9 %
10. Miller 1974	492	3.3 %	1.7-4.9 %
11. Caranasos et al 1974	6063	2.9 %	2.5-3.3 %
12. McKenzie et al 1976	3556	1.9 %	1.5-2.3 %
13. McKinney and Harrison 1976	216	5.6 %	2.4-8.7 %
14. Frisk et al 1977	442	6.8 %	4.4-9.2 %
15. Stewart et al 1980	60	5.0 %	0-10 %
16. Salem et al 1984	41	12.2 %	2.2-22.2 %
17. Lashaman et al 1986	834	4.2 %	2.8-5.6 %
18. Bigby et al 1987	686	6.9 %	4.9-8.9 %
19. Mitchell et al 1988	6546	1.0 %	0.2-1.8 %
20. Col et al 1990	315	16.8 %	12.6-21.0 %
21. Nelson and Talbert 1996	450	5.3 %	3.2-7.4 %
Pooled data	28017	4.7 %	3.1 - 6.2 %

However, the meaning of this pooled result was limited due to a significant heterogeneity between the individual studies: both the fixed effects and random effect tests for heterogeneity were highly significant (both  $p < 0.001$ ,  $I^2 > 90\%$ ). In order to explore the cause for this heterogeneity, the studies of elderly were analyzed separately. The pooled percentages for the elderly (studies 1–3, 6, 16, 17) was 4.8% (95% confidence interval 3.4–5.2%) and for the studies on younger patients (remainder of studies) 3.5% (95% confidence interval 3.1–3.9%). Although the percentage ADEs in elderly patients tended to be larger than that in the younger ( $0.05 < p < 0.1$ ), the overall percentage was not significantly smaller than the percentage of ADEs in the elderly, which suggests that age was not an important cause for heterogeneity in these studies. However, when we assessed the studies for type of research groups, we found out that in the studies with very high percentages, with percentage ADEs from 14.4 to 33.2% (studies 1, 2, 3 and 18) the investigators were clinicians, while all of the other studies had been performed by epidemiologists and pharmacists.

Study	study size	incidence of adverse drug effects	95% confidence intervals
5. Mjordal et al 2002	681	12.0 %	9.6-14.4 %
6. Onder et al 2002	28411	3.4 %	3.2-3.6 %
16. Caamano et al 2005	19070	4.3 %	4.0-4.6 %
Pooled	48161	3.9 %	3.4-4.4 %

Some studies excluded patients with adverse effects due to dosage errors. In order to assess whether negligence of this criterion influenced the overall results, the studies excluding the dosage errors given in the above table (studies 5, 6, and 16) were assessed separately. The pooled percentage in these studies was significantly smaller than that of the remainder of the studies at  $p < 0.01$ , suggesting that the meta-analysis was not entirely robust against the negligence of this criterion. However, like with the main analysis, the meta-analyses of these subgroup studies were heterogeneous (fixed and random test-statistics both  $p < 0.001$ , and  $I^2$ -value  $>90\%$ ).

## 8.13 Example 2, Meta-regression

In order to simultaneously assess the effects of study-magnitude, patients' age, and type-of-research-group meta-regression was, subsequently, performed. After adjustment for patients' age and type-of-research-group the study-magnitude was no significant predictor of study effect anymore. In contrast, the type-of-research-group was the single and highly significant predictor of study result. We also checked the studies for types of recruitment facilities. No differences in percentages ADEs between university and regional hospitals were found. Of the four studies with the largest percentages ADEs two were university, two regional hospitals.

SPSS ([www.spss.com](http://www.spss.com)) multiple linear meta – regression table with study result (percentage ADEs) as dependent variable and study-magnitude, patients' age, and type-of-research-group as independent variables. After adjustment for patients' age and type-of-research-group the study-magnitude is no significant predictor of study effect anymore. In contrast, the type-of-research-group is the single highly-significant predictor of the study result.

Covariate	Unstandardized coefficients		Standardized coefficients		t	sig.
	B	std error	Beta			
Constant	6.92	1.45			4.76	0.000
Study magnitude	-7.7.e-0.05	0.00	-0.071		-0.50	0.62
Patients' age	-1.39	2.89	-0.075		-0.48	0.64
Type research group	18.93	3.36	0.89		5.64	0.000

Dependent variable: study result; std error = standard error; t = t-value;  
sig. = level of significance.

## 8.14 Example 2, Conclusions

In a linear regression controlled for age and study size it was assessed whether pharmacists report adverse drug effects more often than internists do. The difference was very significantly so, the null hypothesis was rejected. The real burden of ADEs in present health care can probably be best assessed by clinicians who have to make a diagnosis and are, subsequently, in charge for starting a treatment. These professionals are by nature of their daily work experienced with adverse effects of medicines.

Multiple meta-regression was used in the example 2. It is convenient for assessing the causes of heterogeneity in the data, and, in addition, for adjusting the effects of multiple confounders, like age and study size in the example given.

## 8.15 Conclusion

In this chapter of two meta-analyses from open evaluation studies, the null hypothesis of no effect could be rejected. In the first example this was: no difference in cardiac risk between patients with and without coronary collaterals. In the second example this was: no difference in adverse drug admission rates between studies of internists versus those of pharmacists. In the first example the effect of collaterals on events was estimated, in the second the effect of investigatortype on events. In the first example collaterals were in each study, in the second the investigatortype was in just one half of the studies.

Controlled long term open evaluation studies will be often applied, if clinical trials are not feasible. Particularly, for investigations of rare events like adverse drug effects or events in rare subgroups, like myocardial infarctions in patients with

collateral coronary arteries, or iatrogenic admissions to hospitals, only open evaluation studies have been published. These studies are prospective, minimizing the risk of recall bias and other flaws of retrospective studies, including risk factor underestimation, because the severest patients never visited the clinic, and of the presence of multiple confounders, because of many differences between the sick and the controls, and other flaws of retrospective observational research.

Open evaluation studies, albeit explorative in nature, gave rise to relevant conclusions. In addition, investigators of any type of study design try and report the most unbiased version of their study, irrespective of its design. This explains, why meta-analyses of studies of any design are possible, as long as there are controls. It is, of course, comprehensible, that studies with lower quality are more at risk of pitfalls like lack of robustness and homogeneity. In this chapter two recently completed meta-analyses of open evaluation studies are reviewed.

## Reference

More background, theoretical and mathematical information of meta-analyses is given in Statistics applied to clinical studies 5th edition, Chaps. 32–34 and 48, Springer Heidelberg Germany, 2012, from the same authors.

# **Chapter 9**

## **Meta-regression**

### **Multiple Regression as an Alternative to Subgroup Analyses**

**Abstract** Regression can be used for the analysis of heterogeneous meta-analyses, for which an overall pooling procedure is pretty meaningless. In this chapter two recently published meta-analyses from the authors are used as examples of exploratory purposes, the assessment of confounding, and of interaction between predictors on the outcome.

#### **9.1 Introduction**

Regression analysis can be used for the analysis of the meta-data from a heterogeneous interventional meta-analysis, for which an overall pooling procedure is pretty meaningless. Purposes include:

1. the search for significant predictors, otherwise called exploratory purpose,
2. the assessment of confounding otherwise called better performance of a subgroup,
3. the assessment of interaction, otherwise called better performance of a subgroup for one treatment.

More information of these purpose are given in Statistics applied to clinical studies, Chap. 15, Springer Heidelberg Germany, 2012, from the same authors. As an additional point, the issue of meta-analysis-of-regression-studies is different from that of regression-studies-of-a-meta-analysis.

#### **9.2 Example 1, Continuous Outcome**

In a 20 study meta-analysis the incidence of adverse drug effects (ADEs) was assessed. The studies were heterogeneous. Twenty studies assessing the incidence of ADEs were meta-analyzed (Atiqi et al.: Int J Clin Pharmacol Ther 2009; 47: 549–56). The studies were very heterogeneous. It was observed, that studies performed by pharmacists (0) produced lower incidences than did studies performed by internists (1). Also the study magnitude and age was considered as possible causes of heterogeneity. The data file is underneath.

variables	dependent %ADEs	independent Study magnitude	independent Clinicians' study	independent Elderly study	Study no	clinician x elderly
21,00	106,00	1,00	1,00	1,00	1	1
14,40	578,00	1,00	1,00	1,00	2	1
30,40	240,00	1,00	1,00	1,00	3	1
6,10	671,00	0,00	0,00	0,00	4	0
12,00	681,00	0,00	0,00	0,00	5	0
3,40	28411,00	1,00	0,00	0,00	6	0
6,60	347,00	0,00	0,00	0,00	7	0
3,30	8601,00	0,00	0,00	0,00	8	0
4,90	915,00	0,00	0,00	0,00	9	0
9,60	156,00	0,00	0,00	0,00	10	0
6,50	4093,00	0,00	0,00	0,00	11	0
6,50	18820,00	0,00	0,00	0,00	12	0
4,10	6383,00	0,00	0,00	0,00	13	0
4,30	2933,00	0,00	0,00	0,00	14	0
3,50	480,00	0,00	0,00	0,00	15	0
4,30	19070,00	1,00	0,00	0,00	16	0
12,60	2169,00	1,00	0,00	0,00	17	0
33,20	2261,00	0,00	1,00	1,00	18	0
5,60	12793,00	0,00	0,00	0,00	19	0
5,10	355,00	0,00	0,00	0,00	20	0

### 9.2.1 Exploratory Purpose

Simple linear regressions with %ADEs as dependent and three predictors as independent variables including

study magnitude

ageclass

investigatortype

was performed using SPSS statistical software. Enter the above data in the software program.

**Then command**

Analyze...Regression...Linear...Dependent: % ADEs ....Independent(s): Study magnitude...click OK.

Repeat the procedure with the other two predictors.

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	11,989	2,296		5,222	,000
study-magnitude	,000	,000	-,355	-1,609	,125

a. Dependent Variable: percentageADEs

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	7,950	2,258		3,521	,002
elderly=1	6,400	4,122	,344	1,553	,138

a. Dependent Variable: percentageADEs

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	6,150	1,101		5,583	,000
clinicians=1	18,600	2,463	,872	7,552	,000

a. Dependent Variable: percentageADEs

The above output sheets show, that only investigatortype was statistically significant.

### 9.2.2 Confounding

In order to explore, what were the independent determinants of hospital admissions due to adverse drug effects, and whether confounding was in the study, multiple linear regressions were performed. Commands similar to the ones in the previous section are given. In the output sheets the underneath table is observed.

Coefficients <sup>a</sup>					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	10,100	2,247		4,495	,000
study-magnitude	-,001	,000	-,466	-2,264	,037
elderly=1	8,515	3,833	,457	2,221	,040

a. Dependent Variable: percentageADEs

The above table shows that the model with two predictors, study magnitude and ageclass, unlike the above simple linear regressions, showed, that both predictors were statistically significant at ,037 and ,040. This would mean that after adjustment for one predictor, the other predictor is now statistically significant, and that one predictor was, thus, a confounder of the other. Next, a multiple linear regression will be performed with percentage ADEs as outcome variable and the study magnitude, the type of investigators (pharmacist or internist), and the age of the study populations as predictors. For analysis the statistical model Linear in the module Regression is required.

### Command

Analyze...Regression...Linear...Dependent: % ADEs ....Independent(s): Study magnitude, Age, and type of investigators....click OK.

### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,879 <sup>a</sup>	,773	,730	4,54549

a. Predictors: (Constant), clinicians=1, study-magnitude, elderly=1

### Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	6,924	1,454		4,762	,000
study-magnitude	-7,674E-5	,000	-,071	-,500	,624
elderly=1	-1,393	2,885	-,075	-,483	,636
clinicians=1	18,932	3,359	,887	5,636	,000

a. Dependent Variable: percentageADEs

The above table is in the output sheets, and shows the results. After adjustment for the age of the study populations and study magnitude, the type of research group was the single and very significant predictor of the heterogeneity. This analysis is, thus, Adjusted for confounding. The b-value (regression coefficient) of investigatortype (clinicians) in the simple linear regression was 18.600. It somewhat rose in the multiple regression adjusted for confounding, and it did so from 18.600 to 18.932 Obviously, internists more often diagnose ADEs than pharmacists do. Indeed, the subgroup of internist produced larger numbers of adverse effects and this subgroup was, thus, a confounder.

### **9.2.3 Interaction**

The above predictor model showed that study magnitude and ageclass were no longer statistically significant, while the investigatortype (clinicians) was very significant. This would mean, that the investigatortype is not only an independent determinant of adverse events, but also, at least partially, a confounder of study magnitude and ageclass. This is pretty unexpected, since in the above simple linear regressions, neither of the two were significant. However, the explanation might very well be a significant interaction variable “study magnitude x ageclass” (see also data file on 1st page of this chapter). When, from the above example, the presence of interaction is assumed, an interaction variable has to be added to the regression model. Subsequently commands similar to the above ones including the interaction variable are given. The underneath output sheets come up after pressing OK.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,922 <sup>a</sup>	,851	,811	3,80615

a. Predictors: (Constant), interaction, study-magnitude,  
elderly=1, clinicians=1

Model	Coefficients <sup>a</sup>				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	6,958	1,218	5,714	,000
	study-magnitude	,000	,000	-1,565	,138
	elderly=1	3,381	2,958	,182	,271
	clinicians=1	26,730	3,961	1,253	,000
	interaction	-15,069	5,389	-,631	,014

a. Dependent Variable: percentageADEs

The correlation coefficient of the multiple linear regression model not accounting interaction was 0.879, with r-square 0.773. After the addition of interaction to the multiple linear regression the correlation coefficient rose considerably, from 0.879 to 0.922. The r-square as a measure for percentage of the outcome predicted by the x-variables (the predictors) rose from 0.773 (77%) to 0.851 (85%). The effect of type of investigator remained statistically very significant. However, the interaction variable between type of investigator and age class appeared to be statistically significant as well. The p-value was 0.014. Obviously, not only was the investigator type a confounder, because the internists diagnosed many more adverse drug effects than the pharmacists did, but also did the subgroup of internists observe significantly more adverse effects in the elderly than it did in the younger patients. An unpaired t-test of the patients with interaction versus those without produced in the no interaction group a mean of 7,7 adverse effects, in the yes interaction group a mean of 21.9 adverse effects.

This result was significantly different at  $p = 0.006$ , and underscored the above regression analysis.

In conclusion, the current example shows that in a heterogeneous meta-analysis of 20 studies only a single of three predictor variables was statistically significant. However, the other two predictors were statistically significant in a two predictor regression model, and the insignificant effects were probably due to a confounding effect of the single significant predictor. Finally a significant interaction between the two initially insignificant predictors on the outcome was observed. The example shows that it does make sense to assess different regression models. This will be particularly relevant, if the models are based on sound theoretical arguments, like those described in the current example.

### 9.3 Example 2, Binary Outcome

Nine studies of risk of infarction of patients with coronary artery disease and collateral coronary arteries were meta-analyzed. The outcome was odds ratio of infarct with versus without collaterals, while the predictor variables were the odds ratios of hypertension and smoking in the patients with versus those without collaterals. This study was published by our group (Akin, Yetgin, Brugts, Dirkali, Zijlstra, Cleophas, Effects of collaterals on deaths and reinfarctions in patients with coronary artery disease, (Neth Heart J 2013; DOI 10.1007). We should add, that, in this study, odds ratios without measure of spread were used as outcome variable. Odds ratios without measure of spread is a pretty much flawed point estimator, and the information of the study sample sizes given in the example cannot compensate for this flaw. But, for now, we will just accept the missing information. The metadata for the meta-analysis are underneath.

Study	OR infarct	OR hypert	OR smoker	sample size	interaction hypert x smoking
1	,44	1,12	,93	1164,00	1,04
2	,62	1,10	,93	42,00	1,02
3	,59	,69	1,85	959,00	1,04
4	,30	,85	,78	712,00	,68
5	,62	,83	1,09	315,00	,90
6	1,17	1,02	1,28	379,00	1,22
7	,30	,17	,27	235,00	,46
8	,70	,79	1,25	4182,00	1,13
9	,26	,74	,83	81,00	,61

OR infarct = ratio of odds of infarct with collaterals versus the odds of infarct without the collaterals

OR hypert = ratio of odds of hypertension with collaterals versus the odds of hypertension without collaterals

OR smoker = ratio of odds of being smoker with collaterals versus the odds of being smoker without collaterals.

Analyses were again performed using SPSS statistical software, the commands are like in the Example 1 of this chapter.

#### 9.3.1 Exploratory Purpose

Explorative regression of the effects of smoking and hypertension on the odds ratio of infarction were not statistically significant, neither in the simple linear regression nor in the multiple regression. The tables with results are below.

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
	B	Std. Error	Beta			
1 (Constant)	,184	,227			,810	,445
ORsmoking	,363	,206	,554		1,760	,122

a. Dependent Variable: ORinfarction

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
	B	Std. Error	Beta			
1 (Constant)	,208	,288			,724	,493
ORhypertension	,427	,336	,433		1,270	,245

a. Dependent Variable: ORinfarction

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
	B	Std. Error	Beta			
1 (Constant)	,044	,304			,146	,889
ORsmoking	,298	,232		,454	1,286	,246
ORhypertension	,254	,349		,257	,728	,494

a. Dependent Variable: ORinfarction

### 9.3.2 Confounding

The effect of hypertension on infarction did not become statistically significant after adjustment for smoking. Neither did the effect of smoking become so after adjustment for hypertension. And, thus, no confounding was in this multiple regression analysis.

### 9.3.3 Interaction

Multiple linear regression with hypertension and smoking, and, in addition, the *interaction* of the two as predictors, showed that only the interaction variable produced a p-value <0.05. The underneath table gives the results of the interaction model.

Coefficients<sup>a</sup>

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
	B	Std. Error	Beta			
1 (Constant)	-,228	,235		-,361	-,971	,376
ORhypertension	-,357	,332		-,247	-1,074	,332
ORsmoking	-,162	,235		-,247	-,690	,521
hypertens xsmoking	1,376	,509	1,260	1,260	2,703	,043

a. Dependent Variable: ORinfarction

With the sample size of the studies included this p-value fell to 0.027. This is shown in the table below.

Coefficients<sup>a</sup>

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
	B	Std. Error	Beta			
1 (Constant)	-,287	,211		-,566	-1,361	,245
ORhypertension	-,559	,322		-,309	-1,736	,158
ORsmoking	-,203	,209		1,605	-,968	,388
hypertens xsmoking	1,752	,512	1,605	1,605	3,421	,027
sample size	-8,170E-005	,000	-,373	-,373	-1,544	,197

a. Dependent Variable: ORinfarction

After removing hypertension and smoking as predictors, the p-value further fell to 0.010 as shown underneath.

Model	Coefficients <sup>a</sup>				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-.294	,230		-1,275	,249
hypertens xsmoking	,987	,264	,904	3,736	,010
sample size	-4,447E-005	,000	-,203	-,839	,433

a. Dependent Variable: ORinfarction

After removing the sample size, the p-value even fell to 0.007, as shown below.

Model	Coefficients <sup>a</sup>				
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-,251	,220		-1,143	,291
hypertens xsmoking	,896	,236	,821	3,801	,007

a. Dependent Variable: ORinfarction

This would mean that in smokers the presence of hypertension produced a very significant increase of the risk of infarction, and likewise in hypertensives the presence of smoking produced a very significant increase of the risk of infarction. These effects were not observed in the above explorative statistics.

In conclusion, the current example shows that in a heterogeneous meta-analysis of 9 studies not a single of two predictor variables was statistically significant, although a trend to statistical significance was observed with smoking at  $p = 0.122$ . This meant that neither confounding was in this regression. However, a regression model including an interaction variable of the two predictors was very significant at  $p = 0.007$ . This remained statistically significant after the addition of other variables including the two initial predictors and the sample size, and, therefore, seems to be a meaningful finding. The example shows that it does make sense, if one regression model does not produce significant effects, to assess another regression model. This will be particularly relevant, if the models are based on sound theoretical arguments, like those described in the current example.

## 9.4 Conclusion

Regression analysis can be used for the analysis of the meta-data from a heterogeneous interventional meta-analysis, for which an overall pooling procedure is pretty meaningless. Purposes include: an exploratory purpose, confounding, interaction.

The current examples show that in heterogeneous meta-analyses confounding and interaction can be easily detected. It does make sense, if one regression model does not produce significant confounding and interaction effects, to assess another regression model. This will be particularly relevant, if the models are based on sound theoretical arguments, like those described in the current examples.

We should add that the meta-analysis of regression studies is not a regression analysis of a meta-analysis. Due to the omnipresent computer the use of arithmetically increasingly complex methods has expanded. The field of statistics has now great difficulty with finding adequate names for its novel methods. Often a name represents various methods like, e.g., the term “mixed model” is used both for mixed effects models (Statistics applied to clinical studies, Chap. 56, Springer Heidelberg Germany, 2012, from the same authors) and mixed linear models (Statistics applied to clinical studies, Chap. 55, Springer Heidelberg Germany, 2012, from the same authors), although the two methods are entirely different. A similar phenomenon is observed with the term “meta-regression”. Although usually applied to name the regression of a meta-analysis, it is currently also sometimes used to name the meta-analysis of regression analyses. Like with meta-analysis of the means and other point estimators of different studies, the main results of regression analyses may be pooled in order to improve the power of testing. In the underneath table an example is given. The B-value (regression coefficient) behaves like a mean value and can be considered to follow a t-distribution. Pooling can be performed as explained in the Chaps. 1 and 2 of this edition. If the magnitudes of the standard deviations or the B-values are very different across studies, a more adequate calculation of the t-values should be performed. It can be observed in the underneath table, that, indeed, the pooled result of the three regression studies produced a much better t- and p-value than did the separate studies.

Study	B	SE	n	dfs	t	p
No. 1	1.5	0.8	20	19	1.875	0.076
2	1.7	0.9	20	19	1.888	0.074
3	1.9	1.0	20	19	1.900	0.073
Pooled result	5.1	1.6	60	57	3.259	0.002

B = regression coefficient, SE = standard error, n = sample size,  
t = t-value, p – value, pooled t-value is calculated according to  

$$(B_1 + B_2 + B_3) / \sqrt{(SE_1^2 + SE_2^2 + SE_3^2)}$$
.

We should also add that, regression analyses of meta-data from studies with multiple categorical rather than continuous outcome and predictors variables provide better sensitivity of testing with random intercept than with fixed intercept regression models. The Chap. 13 of the current edition reviews this subject and gives examples.

## Reference

More information and stepwise analyses of meta-regression models are given in Statistical applied to clinical studies 5th edition, Chap. 34, Springer Heidelberg Germany, 2012, from the same authors.

# Chapter 10

## Meta-analysis of Diagnostic Studies

### Diagnostic Odds Ratios and Summary Receiver Operating Curves

**Abstract** Diagnostic reviews often include the sensitivity/specificity results of individual studies. A problem occurs when these data are pooled, because the correlation between sensitivity and specificity is generally strong negative, causing overestimation of the pooled results. Diagnostic odds ratios may avoid this problem.

This chapter reviews advantages and limitations of diagnostic odds ratios (DORs).

Forty-four previously published diagnostic studies are used as an example. DORs can be readily implemented in diagnostic research. Advantages include that they adjust for the negative and curvilinear correlations between sensitivities and specificities. Limitations include that the outcome parameter is a summary estimate of both sensitivity and specificity, and, that the magnitude of the studies included are not taken into account.

#### 10.1 Introduction

In the past few years many novel diagnostic methods have been developed, including multi-slice computer tomography, magnetic resonance, positive emission tomography and many more methods. Studies evaluating their respective sensitivities and specificities have been published, and meta-analyses of these studies can now be performed in order to establish whether the findings are consistent and can be generalized across populations and morbidity/treatment variations. Sensitivity and specificity are estimators of accuracy of diagnostic methods as explained in the underneath diagram.

Diagnostic test	Gold standard test		positive	negative
	positive	TP	FP	
negative	FN		TN	

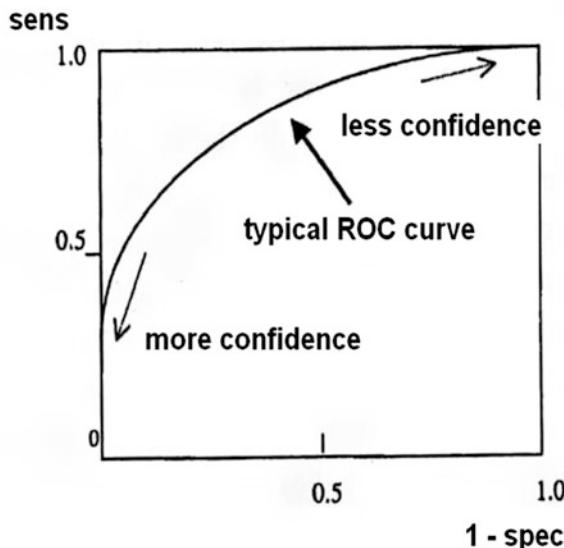
TP = number of true positive, FP of false positive, FN of false negative, and TN of true negative patients in a diagnostic study.

Sensitivity = true positive rate (TPR) =  $TP/(TP + FN)$

Specificity = true negative rate (TNR) =  $TN/(TN + FP)$ .

$$\begin{aligned}1 - \text{Specificity} &= (TN + FP)/(TN + FP) - TN/(TN + FP) \\&= FP / (TN + FP) = \text{FPR} \text{ (false positive rate)}\end{aligned}$$

An intuitive approach to meta-analysis of diagnostic studies is to pool the odds of sensitivity ( $= TPR/(1 - TPR)$ ), and, that of specificity ( $= TNR/(1 - TNR)$ ) of the separate studies. Sensitivities and specificities are, however, dependent on one another, and, in addition, in a non-linear manner as shown in the summary receiver operated characteristic (ROC) curve from the underneath figure. It pictures a summary ROC curve with the proportion of true positive patients drawn against the proportion of false positives. With many diagnostic tests, tests results do not necessarily fall into one of two categories, but rather into categories with more or less confidence in the presence of a disease.



In order to account for these problems Moses and Wittemberg (Stat Med 1993; 12: 1293–316) proposed diagnostic odds ratios of the sensitivities versus the specificities (DORs). In recent years this approach has been increasingly pursued (Hasselblad, Psychol Bull 1995; 117: 167–78, Irwig et al, J Epidemiol 1995; 48: 119–30, Walter SD, Stat Med 2002; 21: 1237–56, Glas et al, J Clin Epidemiol 2003; 56: 1129–35, Bipat et al, Gynecol Oncol 2003; 91: 59–66). The current paper using the results of published diagnostic studies (Scheidler et al, JAMA 1997; 278: 1096–101) as an example, reviews advantages and disadvantages of this novel method and discusses alternative possibilities.

## 10.2 Diagnostic Odds Ratios (DORs)

The accuracy of a diagnostic test is usually summarized by two statistics: the true-positive-rate (TPR) or sensitivity, and the true-negative-rate (TNR) or specificity. They are often used to draw ROC curves. See the above figure. Instead of the dual approach of sensitivity and specificity, accuracy can also be summarized by the diagnostic odds ratio (DOR):

$$\text{DOR} = \frac{\text{sensitivity}/(1 - \text{sensitivity})}{(1 - \text{specificity})/\text{specificity}}$$

The DOR is an interesting term, since it compares the odds of true positive patients with that of false positives, and, thus, summarizes the overall accuracy of a diagnostic test. A problem is, that, like any odds ratio, it does not follow a Gaussian distribution, and a logarithmic transformation is required. The following model is often applied.

$$\ln(\text{DOR}) = \ln\left(\frac{\text{TPR}}{1 - \text{TPR}}\right) - \ln\left(\frac{\text{FPR}}{1 - \text{FPR}}\right) \text{ and}$$

$$S = \ln\left(\frac{\text{TPR}}{1 - \text{TPR}}\right) + \ln\left(\frac{\text{FPR}}{1 - \text{FPR}}\right)$$

where TPR and FPR are the true and false positive rates and  $\ln$  means the natural logarithm. A simple linear regression analysis according to  $\ln(\text{DOR}) = a + b.S$ , with  $\ln(\text{DOR})$  as dependent and  $S$  as independent variable, is used to fit the data. Although this is not obvious from the model as given, this method is often successful in producing a rather close linear fit for the data. The  $a$ -value is the intercept and the  $b$ -value the regression coefficient. If sensitivity equals specificity, then  $\text{TPR} = \text{TNR} = 1 - \text{FPR}$  and  $S$  reduces to 0. And so, the magnitude of the  $\ln(\text{DOR})$  at that point equals the  $a$ -value. The DOR can be calculated by back-log-transforming the calculated intercept.

## 10.3 Example

An example of a meta-analysis of 17 diagnostic studies of lymphangiography for assessment of lymph nodemetastases is given underneath.

$tp$  = true positive,  $fp$  = false positive,  $fn$  = false negative,  $tn$  = true negative patients.

Study No.	tp	fp	fn	tn
1.	0	1	6	17
2.	12	3	3	7
3.	4	1	2	13
4.	10	4	3	25
5.	3	1	4	12
6.	9	3	3	29
7.	20	4	8	31
8.	17	5	7	21
9.	2	0	9	32
10.	3	1	9	38
11.	1	1	2	18
12.	5	2	2	61
13.	21	8	40	184
14.	4	3	9	42
15.	0	0	5	15
16.	7	11	22	158
17.	3	3	2	29

The lymph nodes of the same studies were also analyzed with computerized tomography (CT) and with magnetic resonance imaging (MRI). The underneath tables give the results of the CT (17 studies) and MRI (10 studies) analyses.

Study No.	tp	fp	fn	tn
1.	19	1	10	81
2.	8	9	2	13
3.	41	1	12	49
4.	5	1	2	18
5.	45	58	32	165
6.	8	6	2	32
7.	5	8	1	7
8.	15	17	11	52
9.	16	11	8	24
10.	4	8	2	25
11.	4	12	10	70
12.	10	4	4	55
13.	2	5	6	23
14.	7	10	7	30
15.	4	50	12	135
16.	8	3	1	37
17.	4	3	0	14

Study No.	tp	fp	fn	tn
1.	9	2	2	41
2.	3	6	5	32
3.	3	2	1	16
4.	3	1	12	44
5.	0	0	5	15
6.	7	2	22	167
7.	12	4	4	29
8.	23	5	14	230
9.	8	5	5	53
10.	16	2	2	22

The results of 17 diagnostic studies of imaging techniques for lymph node metastases are used. For example, the ln(DOR) and S-values as calculated from the lymphangiography-studies are entered into the SPSS software program. We command statistics; regression; linear. The program produces an a-value of 2.09 (standard error (SE) = 0.35). The diagnostic odds ratio at the point S = 0 is then found by taking invert natural logarithm of  $2.09 = 8.08$  (SE = 1.35). A summary of the results of the regression analyses are below. The magnitude of DORs at S = 0 can be used to estimate the level of overall accuracy of the diagnostic method. Table 4 shows that MRI imaging is significantly more accurate than the other two methods of cardiac imaging at  $p < 0.001$ .

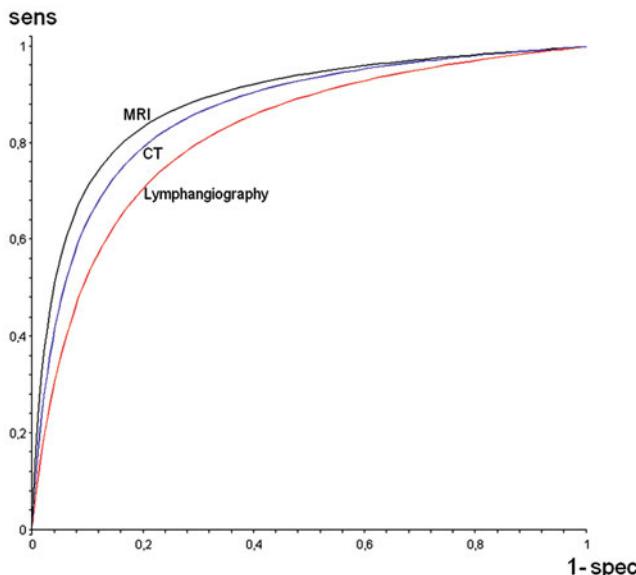
diagnostic modality	intercept (SE) (a-value)	regression(SE) coefficient (b-value)	DOR at S = 0 (SE) (p-values)
1. Lymphangiography	2.09 (0.30)	-0.35 (0.20)	8.08 (1.35) (<0.001 vs CT)
2. CT	2.84 (0.44)	0.23 (0.14)	17.16 (1.55) (<0.001 vs MRI)
3. MRI	3.51 (0.56)	0.25 (0.17)	33.45 (1.75) (<0.001 vs CT)

## 10.4 Constructing Summary ROC Curves

The results of the separate studies are, thus, used to calculate the best fit a and b for the data. Subsequently, the underneath equation is adequate to construct the best fit summary ROC curve from the a- and b-values:

$$\text{TPR} = \left[ 1 + e^{-a(1-b)} \left( \frac{1 - \text{FPR}}{\text{FPR}} \right)^{(1+b)(1-b)} \right]^{-1}$$

We enter the equation into Maple 9.5 software program for making graphs and fill out the a- and b-values. Then the software program produces the best fit ROC curves for the three diagnostic methods.



The above graph gives the summary ROC curves for the three diagnostic modalities. A diagonal line drawn from the top of the y-axis to the right end of the x-axis would contain all points of summary ROC curves where sensitivity equals specificity, and thus  $S = 0$ . Along this diagonal line the distance of the MRI curve to the top of the y-axis would be shorter than that of the other curves, indicating a better accuracy of this diagnostic method. This is supported by a significantly larger DOR at  $p < 0.001$ . Sens = sensitivity, 1-spec = 1- specificity.

The curve closest to the top of the y-axis provides the best overall accuracy. A diagonal line from the top of the y-axis to the right end of the x-axis would contain all points on the summary ROC curves where sensitivity equals specificity, and thus

$S = 0$ . Along this diagonal line the distance from the MRI curve would be shorter than that of the other curves, indicating a better accuracy of this diagnostic method. This is supported by a significantly larger DOR at  $p < 0.001$ .

The distances from the top of the y-axis to the MR/CT/lymphangiography summary ROC curves can be calculated using Pythagoras' equation for rectangular triangles:

$$\begin{aligned}\text{Distance from top of the y-axis} &= \sqrt{[(1-\text{sensitivity})^2 + (1-\text{specificity})^2]}, \\ \text{for the MR curve} &\quad = \sqrt{(0.18^2 + 0.18^2)} = 0.25, \\ \text{for the CT curve} &\quad = \sqrt{(0.22^2 + 0.22^2)} = 0.31, \\ \text{for the lymphangiography curve} &= \sqrt{(0.26^2 + 0.26^2)} = 0.37.\end{aligned}$$

## 10.5 Alternative Methods

Diagnostic odds ratios (DORs) can be readily implemented in diagnostic research. An advantage of the DOR approach to meta-analysis of diagnostic tests is that it accounts the special correlation between sensitivities and specificities of studies included. However, some limitations have to be mentioned. First, as the outcome parameter is a summary estimate of both sensitivity and specificity, no summary estimates of sensitivity or specificity are available. Second, the magnitude of the studies included in the meta-analyses are not taken into account. So, small studies have similar weight compared to large studies.

As an alternative, multivariate methods for pooling the meta-data accounting sensitivities and specificities, can be used. For example, multivariate methods like multivariate analysis of variance (MANOVA) with sensitivity and specificity as outcome variables and different diagnostic modalities as predictor variable can produce results similar to those of the DOR method, and, in addition, produce sensitivities and specificities separately and, at the same time, adjusted for their interaction. A limitation with this approach is, that, again, the magnitude of the separate studies is not accounted, and that the numbers of studies included in the meta-analyses is often too small for reliable testing. A rule of thumb is that at least 10 studies per variable are required for multivariate analyses.

We should add, that, like with therapeutic meta-analyses, it is appropriate to assess diagnostic meta-analyses for the usual pitfalls of meta-analysis including publication bias, clinical heterogeneity, and lack of robustness.

## 10.6 Conclusions

Reported sensitivities and specificities of different studies assessing similar diagnostic tests are not only negatively correlated, but also in a curvilinear manner. It is appropriate to take this negative curvilinear correlation into account in the data

pooling of the meta-analyses of such studies. Diagnostic odds ratios can be applied for that purpose.

## References<sup>1</sup>

- Glas AS, Roos D, Deutkom M, Zwinderman AH, Bossuyt PM (2003) Tumor markers in the diagnosis of primary bladder cancer, a systematic review. *J Urol* 169:1975
- Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH et al (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 58:982–990
- Van Houwelingen HC, Zwinderman AH (1993) A bivariate approach to meta-analysis. *Stat Med* 12:2273–2284

---

<sup>1</sup>Three more diagnostic meta-analyses have been performed by members of our group.

# Chapter 11

## Meta-Meta-analysis

### A Shift from a Single to Multiple Meta-analyses

**Abstract** A meta-meta-analysis is a meta-analysis of multiple meta-analyses. In this chapter a meta-meta-analysis from the authors is performed for the purpose of re-assessment of the pitfalls of the original meta-analyses with increased power and sample size. Also a meta-meta-analysis, performed for meta-learning purposes by Juni et al., and published in the BMJ of 2001, is reviewed.

#### 11.1 Introduction

Today numerous meta-analyses have been conducted. A shift from single meta-analyses to multiple meta-analyses is one of the consequences. First of all, the results of original meta-analyses are sometimes combined for the purpose of a re-assessment of the pitfalls of the original meta-analyses with increased power. Second, meta-analyses of meta-analyses are sometimes performed for meta-learning purposes. Meta-analytic thinking and meta-learning are new fields. Kate (Cross-disciplinary perspective on meta-learning for algorithm selection, ACM Computing Surveys 2009; 41: doi 10.1145) gives some examples of novel methodologies and algorithms, including meta-cognition, meta-knowledge, higher order of thinking, meta-learning, meta-strategic knowledge, awareness of learning processes rather than knowledge and thinking skills. And in the medical world the concept of meta-analytic thinking has begun (World Heritage Encyclopedia, Meta-Analytic Thinking, Copyright © 2016 World Public Library). More information of meta-learning is given in the Chap. 25.

In this chapter we will describe examples of (1) a meta-meta-analysis for the purpose of re-assessment of the pitfalls of the original meta-analyses with increased power, (2) a meta-meta-analysis for meta-learning purposes.

## 11.2 Example 1, Meta-Meta-analysis for Re-assessment of the Pitfalls of the Original Meta-analyses

A totally different example is given underneath. A meta-analysis of two meta-analyses, otherwise called meta-meta-analysis or meta-epidemiological study, is reviewed. The traditional publication bias, heterogeneity, and robustness pitfalls remained present, but the larger number of studies enabled to point out several possible causes of heterogeneities, for example the presence of outlier studies, and of possibly double publications and of the duplicate publication effect of separate male and female studies from a single population.

[Am J Cardiol.](#) 2000 Nov 1;86(9):1005-9, A8.

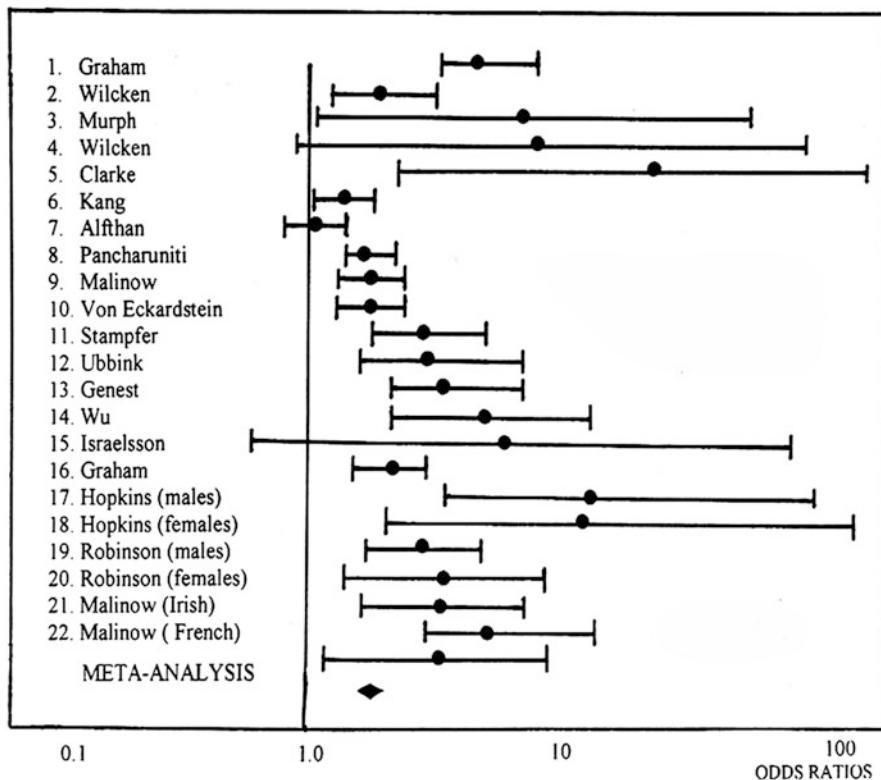
**Homocysteine, a risk factor for coronary artery disease or not?**

[Cleophas TJ<sup>1</sup>, Hornstra N, van Hoogstraten B, van der Meulen J.](#)

### Abstract

This meta-analysis suggests that homocysteine may not be as harmful for the heart as it seems. At the same time, however, homocysteine may be an indicator for unhealthy lifestyles, and therefore, an important variable for cardiologists to take into account when assessing coronary artery disease.

An important argument in favor of the concept that homocystinemia is not very harmful to the heart was given by a study of our group Boers et al, [N Engl J Med](#) 1985; 313: 709-15: slightly or even moderately elevated levels as seen with heterozygous cystathione deficiency did not give rise to increased risk of coronary artery disease (CAD), and that even the homozygous form of this disease with more than 100-fold levels of plasma homocysteine although it commonly gave rise to severe peripheral vascular disease, gave rise to CAD in only 10 of 629 patients (less than 2% of the cases), which is not different from incidences of CAD in otherwise healthy populations. In contrast, many observational studies (see forest plot underneath) found a significant association. However, problems with observational studies are numerous, including selection bias, placebo effects, confounding, time effects, interactions, etc. Two earlier meta-analyses from our group, one of 22 case-control studies, and one of 11 cohort studies, both provided strong support for the causal factor theory of homocysteine for CAD. The underneath graphs 1 and 2 respectively give summary results and forest plots of the individual studies. Particularly the results of the first meta-analysis were pretty heterogeneous, but the heterogeneous studies were small, and the pooled effect size had a pretty narrow 95% confidence interval, and none of the studies showed mean effect sizes smaller than 1.0.



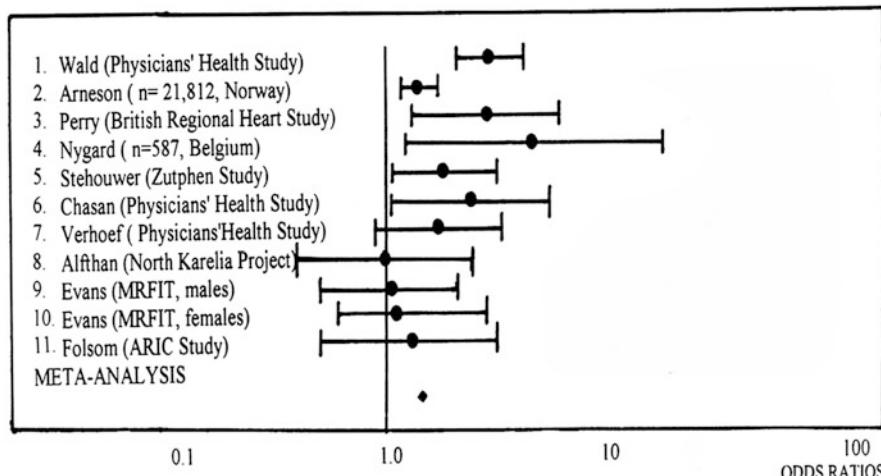
1st Meta-analysis

22 case-control studies

pooled odds ratio 1.62

95% confidence interval 1.50–1.74

$p < 0.001$  versus an odds ratio of 1.0



## 2nd Meta-analysis

11 cohort studies

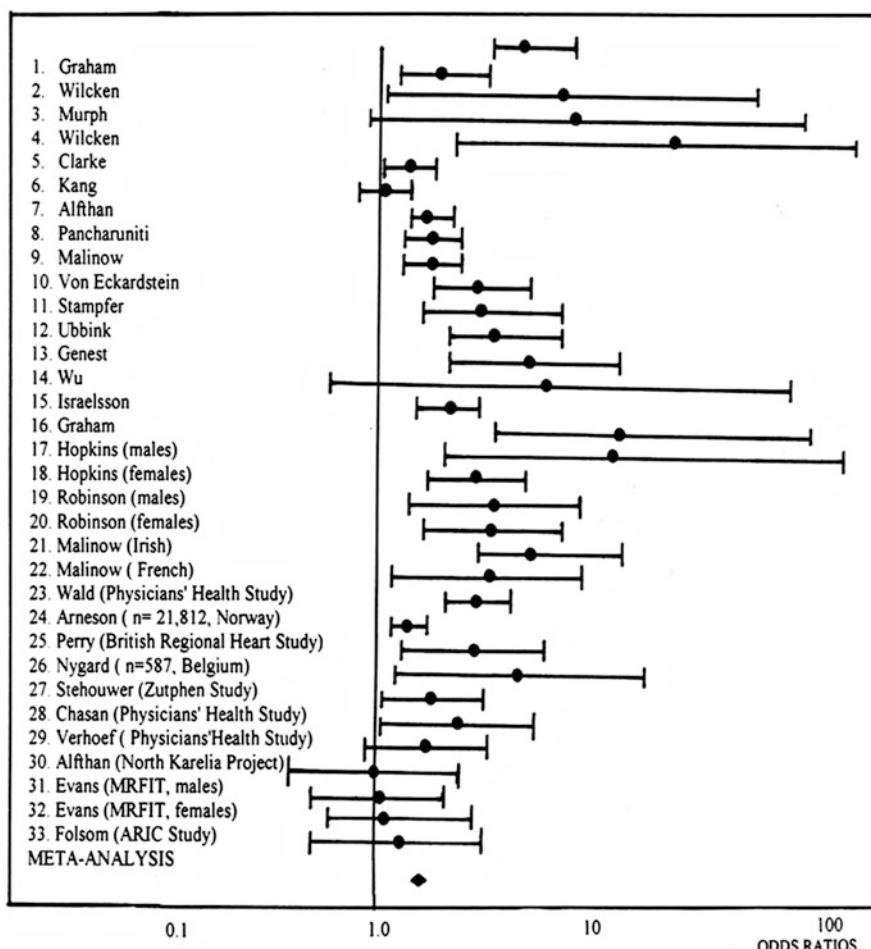
pooled relative risk 1.49

95% confidence interval 1.33–1.67

 $p < 0.001$  versus an odds ratio of 1.0.

The pooled results of the above forest plots from the two meta-analyses are given. A combined analysis of the above two meta-analyses is given next.

In order to enable to better provide better power for demonstrating possible biases, a combined analysis of the above two meta-analyses was performed (Cleophas et al, Am J Cardiol 2004; 86: 1005–9). All of the 33 studies were published between 1976 and 1999, and they involved 16,097 patients. For the meta-meta-analysis the data were assessed for publication bias, heterogeneity and robustness according to the standard guidelines of Oxmann and Guyatt (Guidelines for reading literature reviews, Can Med Assoc J 1988; 138: 697–703). Relative risk were used as a measure of the relation between elevated homocysteine levels and fatal or nonfatal coronary artery disease. For case-control studies the relative odds ratios (ORs) were used as a surrogate measure of the corresponding relative risk (RR).



The null hypothesis, that the pooled OR was not different from 1.0 was tested using chi-square test of added point estimators.

#### Combined Meta-analysis

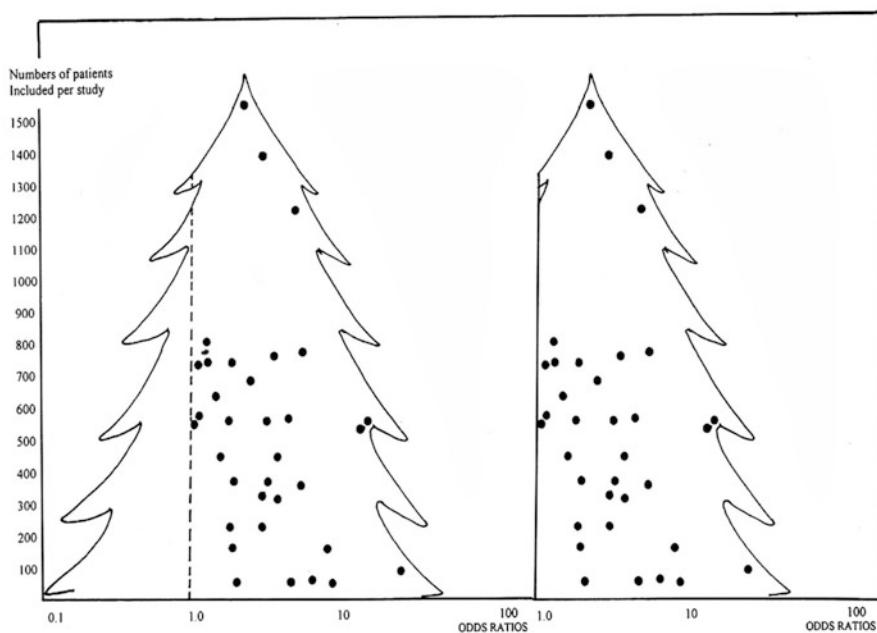
33 studies

pooled relative risk 1.58

95% confidence interval 1.49–1.68

$p < 0.001$  versus an odds ratio of 1.0.

The pitfalls of the meta-meta-analysis were re-assessed next.



The above funnel plot (Christmas tree) graph suggests the presence of considerable publication bias. A Gaussian pattern of the study results was plausible, but smaller studies with smaller results were not published. The presence of publication bias was confirmed by separate pooling of larger and the smaller studies.

#### Small studies

sample size  $< 500$

pooled relative risk 1.88

95% confidence interval 1.72–2.05

$p < 0.001$  versus the large studies.

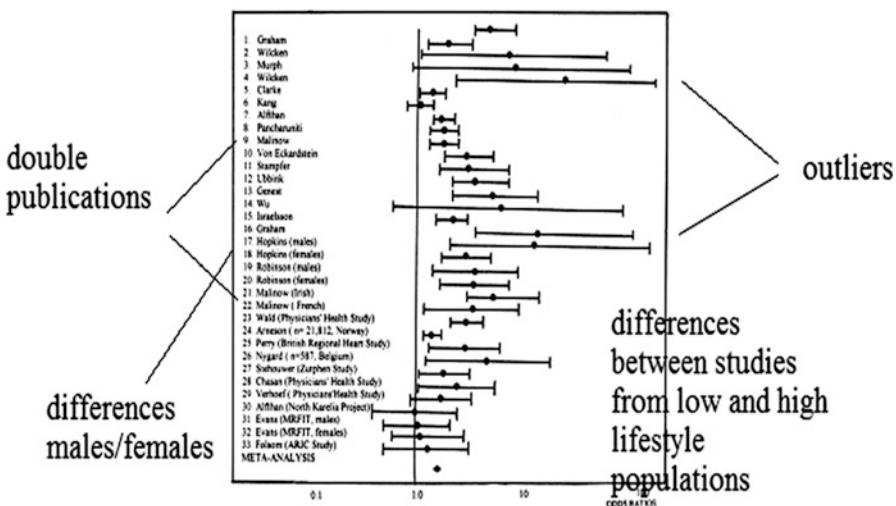
Heterogeneity of main effects (the RRs) between the 33 separate studies was tested using multiple group chi-square test (fixed effect model).

#### Heterogeneity between study's main effects

chi-square  $> 140.0$

$p < 0.0001$

The subsequent random effect analysis produced a lower chi-square test statistic (of only 70.0 instead of 140.0), but the p-value remained statistically significant at 0.001.



Searching for the cause of heterogeneity may be the most important part of the whole enterprise of the current assessment. In the above graph several possible causes could be pointed out.

Suspected causes of heterogeneity between studies were:

1. Outlier studies: the studies 3, 4, 5, 14, 16, 17, are pretty much outlier studies.
2. Authors producing two meta-analyses (are they partly or entirely duplicate publications?).
3. Many authors produce separate meta-analyses for genders (are these studies independent of one another?).
4. Lifestyle is a recognized confounder of deleterious effects of homocysteine. The Bloomberg ranking of countries with healthiest lifestyle were tested against the effect sizes of the studies. No relationship was found.
5. The confidence intervals were, particularly those of the case-control studies were often very wide, and more so than they were in the cohort studies. This is as expected since in case-control studies many differences exist between the sick and the controls).
6. The American studies (6, 8, 9, 11, 28–33) tended to have smaller results than the studies from other countries. Many explanations can be given.

Sensitivity analysis was aimed at demonstrating, whether lower quality studies had a more spectacular result. For that purpose the 22 case-control studies were tested against the 11 cohort studies, using an unpaired t-test.

Pooled effect case-control studies  $1.62 \pm 0.12$

( $0.12 = \text{standard error}$ ),

pooled effects cohort studies  $1.49 \pm 0.17$ ,

t-test

$$t = (1.62 - 1.49) / \sqrt{(0.12^2 + 0.17^2)} = 0.33$$

t smaller than 1.96, not statistically significant.

In spite of a larger pooled effect of the case-control studies, than that of the cohort studies, the difference was not statistically significant. Lack of sensitivity was, thus, not demonstrated.

### 11.3 Example 2, Meta-Meta-analysis for Meta-learning Purposes

**Education And Debate** Systematic reviews in health care

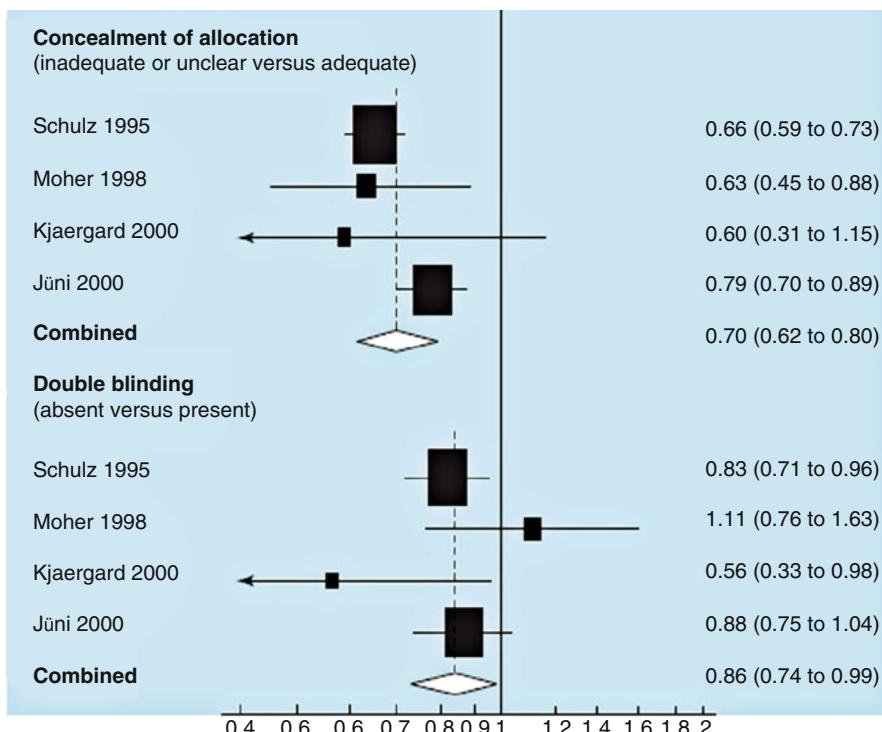
#### Assessing the quality of controlled clinical trials

BMJ 2001; 323 doi: <http://dx.doi.org/10.1136/bmj.323.7303.42> (Published 07 July 2001) Cite this as: BMJ 2001;323:42

As an example of a meta-meta-analysis the systematic reviews in health care article of Juni is given (BMJ 2001; 323: 42–6). The article could be classified as a study characteristics study, and as a study of higher order of thinking (see also Chap. 25, Meta-analytic thinking and other spin-offs of meta-analysis), because, unlike traditional meta-analyses, it did not have the same prior hypothesis, as, that of the primary studies of the meta-analysis, but, rather, a novel post hoc hypothesis, i.e., the hypothesis, that the studies with adequate blinding will have a less impressive result than that of the inadequately blinded studies. A novel meta-analysis of four original meta-analyses, each of them assessing quality criteria of controlled clinical trials, including

- (1) adequacy of allocation sequence,
- (2) adequacy of allocation concealment, and
- (3) adequacy of blinding,

was performed. The meta-analysis of Schultz, the first of four, as shown in the underneath graph, consisted of 250 studies, which had originally already been published in the form of 33 meta-analyses. The main results were summarized in the underneath forest plot.



Forest plots are more profoundly explained in the Chap. 25. In the above graph (copied from the public domain) the black squares present the ratios of the pooled odds ratios of the inadequate studies and the pooled odds ratios of the adequate studies (with the size of the squares adjusted for the numbers of studies included). The odds ratios and the pooled odds ratios can be considered as the treatment effects of the studies and the meta-analyses. Obviously, most of the black squares are on the left side of the odds ratio = 1 ordinate, suggesting, that, generally, the inadequate studies produced better (= smaller) odds ratios than did the adequate studies. The article did not provide data of validity-, quality- assessments, but the overwhelming body of studies included were definitely impressive, and the trends observed were definitely in agreement with the authors' prior hypothesis.

## 11.4 Conclusion

Meta-analysis is no longer a novelty, and numerous meta-analyses have been conducted and published. A shift from single meta-analyses to the meta-analyses of multiple meta-analyses is a consequence. An example is given of a recently performed meta-analyses of two original meta-analyses results for the purpose of

re-assessment of the pitfalls with probably more power. The traditional pitfalls of publication bias, and heterogeneity, did not disappear and the robustness pitfall remained unchanged, but the larger number of studies enabled to point out several possible causes of heterogeneities, for example the presence of outlier studies, and of possibly double publications and of the duplicate publication effect of separate male and female studies. Also an example is given of a meta-analysis of four original meta-analyses assessing the quality of controlled clinical trials, a sort of meta-learning exercise, successfully supporting the authors' prior belief about the effect of flawed double-blinding.

Meta-analysis is no longer a novelty in medicine. Numerous meta-analyses have been conducted for the same medical topic, and there is a trend to meta-analytical thinking (Meta-analysis, Wikipedia 2016, April 11) leading to a shift from single meta-analyses to multiple meta-analyses. The result of different meta-analyses are combined, and the new type of study is called a meta-epidemiological study (for example, Bae, Epidemiol Health 2014; 36: e2014019). The assessment of biases is the main objective, and that of the pooled effects is less so.

## Reference

More background, theoretical and mathematical information of meta-analyses is given in Statistics applied to clinical studies 5th edition, Chaps. 32–34 and 48, Springer Heidelberg Germany, 2012, from the same authors.

# **Chapter 12**

## **Network Meta-analysis**

### **Bayesian Networks, Frequentists' Networks**

**Abstract** Network meta-analysis assesses direct and indirect relationships across a network of predictor and outcome variables. Three previously published meta-analyses assessing the effect of various predictors on the frequency of iatrogenic hospital admissions were re-analyzed separately and in combination. Five different networks were produced by the Konstanz Information Miner (Knime). One network provided a much better fit for the data than the other.

#### **12.1 Introduction**

Bayesian network meta-analysis is able to account not only direct but also indirect relationships across a network of variables. Unfortunately, p-values are not provided, but Akaike information indexes can be used to assess and compare the goodness of data fit of one Bayesian network model against any other Bayesian model. In this chapter the results of regression analyses of meta-data were compared with those of Bayesian networks of the same meta-analyses. It is shown that worthwhile additional information is given by the network methodology, and that one network provides a much better fit for the meta-data than the other does.

## 12.2 Example 1, Lazarou-1 (JAMA 1998; 279: 1200–5)

year of study	investigator type	study size	% admissions due to adverse effects
1995,00	2,00	379,00	5,30
1995,00	2,00	4031,00	4,40
1994,00	1,00	1024,00	10,30
1993,00	2,00	420,00	3,60
1981,00	1,00	815,00	14,80
1979,00	2,00	1669,00	16,80
1977,00	2,00	152,00	7,20
1977,00	1,00	334,00	10,20
1973,00	1,00	11526,00	22,50
1973,00	2,00	658,00	12,20
1971,00	2,00	8291,00	1,20
1970,00	1,00	939,00	10,50
1968,00	1,00	830,00	24,10
1967,00	1,00	267,00	10,90
1966,00	1,00	714,00	13,60
1966,00	1,00	900,00	10,80
1965,00	1,00	500,00	8,20
1964,00	1,00	1014,00	10,20

In the above 18 study meta-analysis, of studies published 1964–1995, the percentage of hospital admission due severe adverse drug effects were assessed. The percentage of adverse drug admissions out of all of the admission was the main outcome measure. As this meta-analysis was very heterogeneous, a multiple linear regression was performed to find out what predictors may have influenced the outcome.

Coefficients<sup>a</sup>

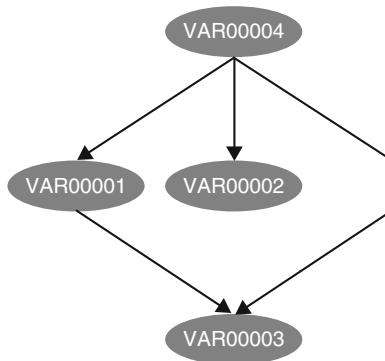
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std.Error	Beta		
1 (Constant)	155,540	280,645		,554	,588
year of study	-,070	,143	-,128	-,487	,634
department	-5,418	3,148	-,453	-1,721	,107
study size	,000	,000	,172	,765	,457

a. Dependent Variable: % severe

This multiple linear regression with percentage severe adverse events as outcome and three predictors, (1) year of study, (2) department (internal medicine or otherwise), and (3) study size, shows, that, with  $p = 0.15$  as cut-off for statistical

significance only the department (= type of investigator, internist or otherwise) is a significant predictor of percentage of adverse effects.

Next, we will perform a Bayesian network assessment, bottom-up reasoning, and we will use for the purpose the Knime and Weka software programs. In order for readers to perform their own analyses examples of step by step analyses are given in Machine learning in medicine a complete overview, the Chaps. 7, 70, 71, 74, Springer Heidelberg Germany, 2015, from the same authors.



#### BOTTOM-UP

var = variable

var 0004 = % severe adverse drug reactions (outcome)

var 0003 = study size

var 0002 = department (1 medicine, 2 = otherwise)

var 0001 = year of publication of report

The percentage adverse reactions (0004) is predicted by (0001) the year of publication of report, (0002) the department, and (0003) the study size directly, and, indirectly, through the year of publication of study. This result, similarly to the above linear regression model, suggests a relationship between hospital admissions due to adverse drug reactions and the type of investigator (here called departments, either internal medicine or otherwise), but the Bayesian network, additionally, suggests, that both year of publication and study size have some predictive meaning.

### 12.3 Example 2, Atiqi (Int J Clin Pharmacol Ther 2009; 47: 549–56)

%ADEs	Study magnitude	Clinicians' study yes =1	Elderly study yes =1	Study no
21,00	106,00	1,00	1,00	1
14,40	578,00	1,00	1,00	2
30,40	240,00	1,00	1,00	3
6,10	671,00	0,00	0,00	4
12,00	681,00	0,00	0,00	5
3,40	28,411,00	1,00	0,00	6
6,60	347,00	0,00	0,00	7
3,30	8601,00	0,00	0,00	8
4,90	915,00	0,00	0,00	9
9,60	156,00	0,00	0,00	10
6,50	4093,00	0,00	0,00	11
6,50	18,820,00	0,00	0,00	12
4,10	6383,00	0,00	0,00	13
4,30	2933,00	0,00	0,00	14
3,50	480,00	0,00	0,00	15
4,30	19,070,00	1,00	0,00	16
12,60	2169,00	1,00	0,00	17
33,20	2261,00	0,00	1,00	18
5,60	12,793,00	0,00	0,00	19
5,10	355,00	0,00	0,00	20

Twenty studies assessing the incidence of ADEs (adverse drug effects) were published 1995–2009, and were meta-analyzed by Atiqi et al. (Int J Clin Pharmacol Ther 2009; 47: 549–56). These studies were also very heterogeneous. It was observed, that studies performed by pharmacists (0) produced lower incidences of ADEs than did the studies performed by clinicians (1). Also, the study magnitude and age of study populations per study were considered as possible causes of heterogeneity. The data are in the above table.

A multiple linear regression will be performed with percentage ADEs as outcome variable and the study magnitude, the type of investigators (pharmacist or internist), and the age of the study populations as predictors. For analysis SPSS statistical software was used. The statistical model Linear in the module Regression is required.

First enter the data in the Data View. Then

#### Command

Analyze...Regression...Linear...Dependent: % ADEs ...Independent(s): Study magnitude, Age, and type of investigators...click OK.

Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	6,924	1,454		4,762	,000
study-magnitude	-7,674E-5	,000	-,071	-,500	,624
elderly=1	-1,393	2,885	-,075	-,483	,636
clinicians=1	18,932	3,359	,887	5,636	,000

a. Dependent Variable: percentageADEs

The above table is in the output sheets, and shows the results. After adjustment for the age of the study populations and study magnitude, the type of research group was the single and very significant predictor of the heterogeneity. Obviously, internists more often diagnosed ADEs than pharmacists did.

In a multiple linear regression with “the % admissions to hospital due to adverse drug effects” as outcome variable, the type of investigator was the only significant predictor at  $p < 0.000$ . The magnitude of the studies and the age of the participants were not significant predictors.

In conclusion, the outcome “the % admissions to hospital due to adverse drug effects” was predicted by

- (1) type of investigator,  $p = 0.0001$
- (2) magnitude of study,  $p = \text{NS!}$
- (3) age of participants,  $p = \text{NS!}$

The Network Meta-Analysis on the Konstanz Information Miner (Knime) and Weka 3.6 software for windows (data mining statistical software from the University of Waikato (New Zealand)) was, subsequently, used to produce a Bayesian network. Bayesian networks, otherwise called structural equation models (SEMs). SEMs are probabilistic graphical models, where regression coefficients are turned into standardized regression coefficients, that can be, simply, added up in order to form DAGs (directed acyclic graphs). If two variables correlate with one another, chance rather than causality may be responsible. If more than two variables correlate the causality is much more probable. In the underneath 4 variable network 4 correlations have been suggesting to support the existence of causal relationships, rather than chance findings. The procedure is called a Bayesian network meta-analysis. It shows, similarly to the multiple linear regression model, that the % of adverse drugs effects is predicted by the type of investigator. However, in addition, a very interesting network of relationships is established. Three more possibly causal relationships are being demonstrated.

1. The type of investigator predicts the age of the participants.
2. The type of investigator predicts the study size, through both direct and indirect effects.
3. The % of adverse drug effect predicts the study size.

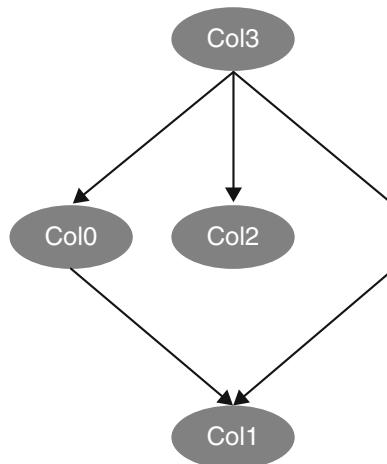
We should add that, unlike the first example, the example shows a *predictive support or top-down reasoning*. The first example showed a *diagnostic support or bottom-up reasoning*. Bayesian networks assess linear or loglinear relationships, and use for the purpose both top-down and bottom-up reasoning, the causal relationships of which is sometimes explained by the parent child relationships. The parent influences the child, top-down reasoning, while the child is dependent on the parent bottom-up reasoning. The approach is somewhat similar to that of discriminant analysis, where dependent variables of a regression model are turned into kind of independent variables, simply, because we more are interested in the independent variables influencing the dependent ones.

col. 3 = column 3 = type of investigator (1 = clinician, 0 = pharmacist)

col. 2 = column 2 = age class (1 = elderly, 0 = younger)

col. 1 = column 1 = study size

col. 0 = column 0 = the % admissions to hospital due to adverse drug effects



### TOP-DOWN

The above graph shows that the Knime and Weka programs have chosen a top-down approach here for explaining what is going on. Again, just like with the network from the example 1, the conclusion can be: the type of investigator is an important predictor of the percentage of hospital admission due to adverse effects. And so, this example of more recent data leads to a conclusion similar to that of the first example.

The Bayesian network does not provide p-values. However, goodness of fit of the network can be assessed with Akaike information index (AIC). The smaller the index, the better the network model fits the data. The Lazarou-1 network produces

an AIC of  $-10977.916\ldots$ , while the Atiqi network does so of  $-2080.435$ . Both AICs are pretty small, and the Lazarou-1 fits even slightly better than the Atiqi network does.

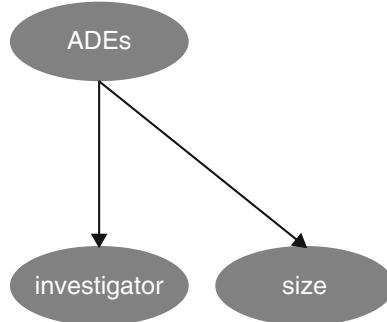
## 12.4 Example 3, Lazarou-1 and Atiqi (JAMA 1998; 279: 1200–5, and Int J Clin Pharmacol Ther 2009; 47: 549–56)

Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	25,985	3,262		7,965	,000
department	-9,514	2,023	-,629	-4,702	,000
study size	-8,906E-005	,000	-,076	-,569	,573

a. Dependent Variable: % severe

The results of two heterogeneous meta-analyses obtained from the meta-analyses from examples 1 and 2 will now be combined in order to provide additional power, and, possibly, new and so far unobserved relationships. The above linear regression table shows that the type of investigator, clinician or pharmacist, is a very significant predictor of the outcome “% severe”, meaning the % admissions to hospital admissions due to adverse drug effects. The b-value is  $-6.29$  with a t-value of  $-4.702$ . This is very good, but better power than the above two tests is not provided. The graph underneath shows the Bayesian network. The software has chosen again a *bottom-up* model underscoring that the type of investigator is an independent predictor of percentages patients hospitalized for adverse drug admissions.



BOTTOM-UP

## 12.5 Example 4, Lazarou-1 and -2 (JAMA 1998; 279: 1200–5)

year of study	investigator type	study size	% admissions due to adverse effects
1995,00	2,00	379,00	5,30
1995,00	2,00	4031,00	4,40
1994,00	1,00	1024,00	10,30
1993,00	2,00	420,00	3,60
1981,00	1,00	815,00	14,80
1979,00	2,00	1669,00	16,80
1977,00	2,00	152,00	7,20
1977,00	1,00	334,00	10,20
1973,00	1,00	11526,00	22,50
1973,00	2,00	658,00	12,20
1971,00	2,00	8291,00	1,20
1970,00	1,00	939,00	10,50
1968,00	1,00	830,00	24,10
1967,00	1,00	267,00	10,90
1966,00	1,00	714,00	13,60
1966,00	1,00	900,00	10,80
1965,00	1,00	500,00	8,20
1964,00	1,00	1014,00	10,20
1996,00	2,00	450,00	5,30
1990,00	1,00	315,00	16,80
1988,00	2,00	6546,00	1,00
1987,00	1,00	686,00	6,90
1986,00	1,00	834,00	4,20
1984,00	2,00	41,00	12,20
1980,00	2,00	60,00	5,00
1977,00	2,00	442,00	6,80
1976,00	1,00	216,00	5,60
1976,00	2,00	3556,00	1,00
1974,00	1,00	6063,00	2,90
1974,00	1,00	492,00	3,30
1874,00	1,00	555,00	1,80
1974,00	1,00	1025,00	3,00
1974,00	1,00	1193,00	5,60
1974,00	1,00	2065,00	2,00
1973,00	2,00	658,00	2,90
1970,00	1,00	939,00	5,10
1967,00	1,00	267,00	4,50
1966,00	1,00	714,00	3,90
1966,00	1,00	900,00	1,70

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-95,518	101,603		-,940	,354
year of study	,054	,052	,180	1,048	,302
department	-3,247	2,040	-,275	-1,591	,121
study size	,000	,000	,057	,346	,732

a. Dependent Variable: % severe

Lazarou (JAMA 1998; 279: 1200–5) provided an additional set of studies providing information about studies who assessed adverse drug reactions in patients already admitted to hospital. The above table provides a linear regression of both the patients admitted because of adverse drug reactions, and those already in hospital with adverse drug reactions. They included 39 studies (Lazarou JAMA 1998; 279: 1202) of percentages of hospitalized patients with adverse drug reactions.

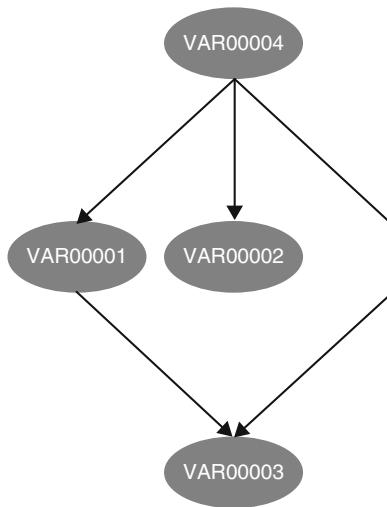
Var 0001 = year of publication,

Var 0002 = investigator (1 = internist, 2 = otherwise),

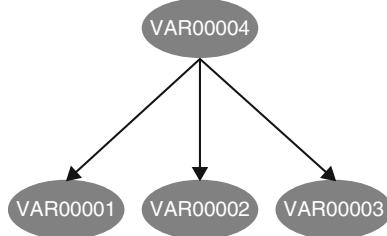
Var 0003 = sample size of study,

Var 0004 = percentage of hospitalized patients for adverse drug reactions.

The above table shows that none of the predictors of % severe (percentage of patients in hospital with adverse drug reactions) were statistically significant, a pretty disappointing result. However, a *bottom-up* network meta-analysis model could be produced using again the above software. It suggested, that investigator type predicted the percentage of patients hospitalized for adverse drug reactions, and was, thus, again entirely in agreement with the results of the above three networks. We should add that, although p-values were not provided by the software, the consistency of the results was, thus, excellent. If you don't have additional data files, another way of assessing consistency of the results of the network meta-analysis is to do some kind of cross validation to assess the reliability of the network.



BOTTOM-UP



BOTTOM-UP

The above network based on the same data as those of the former network is different from the former. Which of the two networks is better. Akaike information index is often used for assessment. It is a goodness of fit test, calculated from the difference between the numbers of parameters minus their likelihood score. The smaller the Akaike information index, the better the fit of the model.

The Akaike information index for the two models are

–6652.384. for the former

and

–37.662... for the latter.

This would mean that the first of the two models provided a much better fit for the data given than the latter did.

## 12.6 Conclusion

Bayesian network meta-analysis is able to account not only direct but also indirect relationships across a network of variables. Unfortunately, p-values are not provided because it works with added standardized regression coefficients for which p-values cannot be obtained, but Akaike information indexes can be used to assess and compare the goodness of data fit of one Bayesian network model versus others. In this chapter the results of regression analyses of meta-data were compared with those of Bayesian networks of the same meta-analyses. It is shown, that worthwhile additional information is given by the network methodology, and that one network provides a much better fit for the meta-data than the others do.

In the past few years the network methodology has also been used as a method for synthesizing information from a network of trials addressing the same question but involving different interventions (Mills et al, JAMA 2012; 308: 1246–53, and Cipriani et al, Ann Intern Med 2013; 159: 130–7). The key assumption with any type of network meta-analysis is, of course, the exchangeability assumption: patient and study characteristics of studies in the meta-analysis must be similar enough to be comparable. The exchangeability assumption can sometimes be tested, but this is virtually always hard to do. E.g., a trial with three treatment arms A versus B versus C, should have the same effects of A versus C and B versus C as separate trials of two treatment arms should have. Currently, Extended Bucher networks (Song et al, BMJ 2011; 343: d4909) and Lumley networks (Stat Med 2002; 21: 2313–24) can cover for more complex networks with contrast coefficients adding up to one (see also the Chap. 22 of the current edition for a review of contrast coefficient meta-analysis).

## Reference

Bayesian networks are reviewed in the Chap. 70, in: Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors.

# Chapter 13

## Random Intercepts Meta-analysis

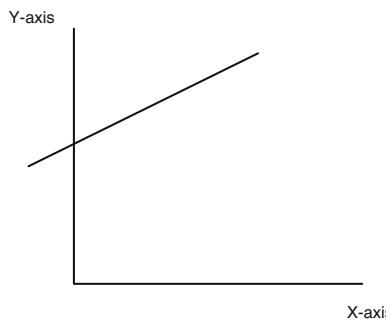
### Multiple Categorical Outcome and Predictor Variables

**Abstract** Meta-analyses with multiple categorical outcome and predictor variables provide better sensitivity of testing with random intercepts than with fixed intercepts models.

Three studies each of them from a different hospital department assessed the effect of ageclass and hospital department on the fall out of bed risk. The random intercept analysis of variance model provided better statistics than the fixed intercept analysis of variance model did.

#### 13.1 Introduction

Meta-analyses with multiple categorical outcome and predictor variables provide better sensitivity of testing with random intercepts than with fixed intercepts models.



Simple linear regression uses equation  $y = a + bx$ ,

Use the equation to make predictions

if we fill out  $x$ -value = 0 => then the equation turns into  $y = a$

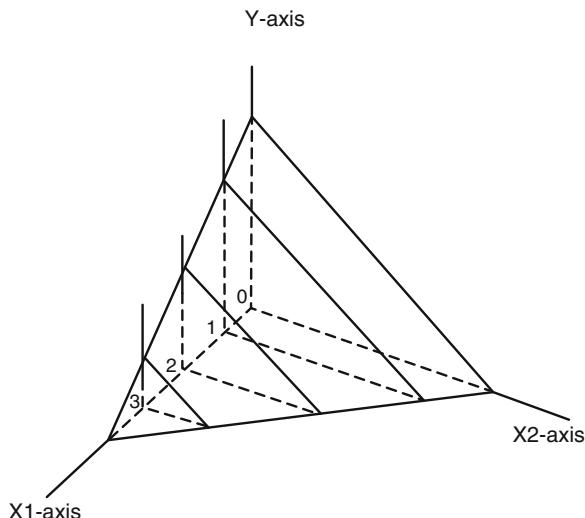
$$x = 1 =>$$

$$y = a + b$$

$$x = 2 =>$$

$$y = a + 2b$$

For each  $x$ -value the equation gives best predictable  $y$ -value, all  $y$ -values constitute a regression line = the best fit for the data (with the shortest distance to the  $y$ -values). In this model only one  $a$ -value, otherwise called the intercept, exists.



Multiple regression with 3 variables uses the equation  $y = a + b_1x_1 + b_2x_2$ .

We can use a 3-axes-model with  $y$ -axis,  $x_1$ -axis and  $x_2$ -axis:

If we fill out  $x_1 = 0$ , then the equation turns into  $y = a + b_2x_2$

$$x_1 = 1$$

$$y = a + b_1 + b_2x_2$$

$$x_1 = 2$$

$$y = a + 2b_1 + b_2x_2$$

$$x_1 = 3$$

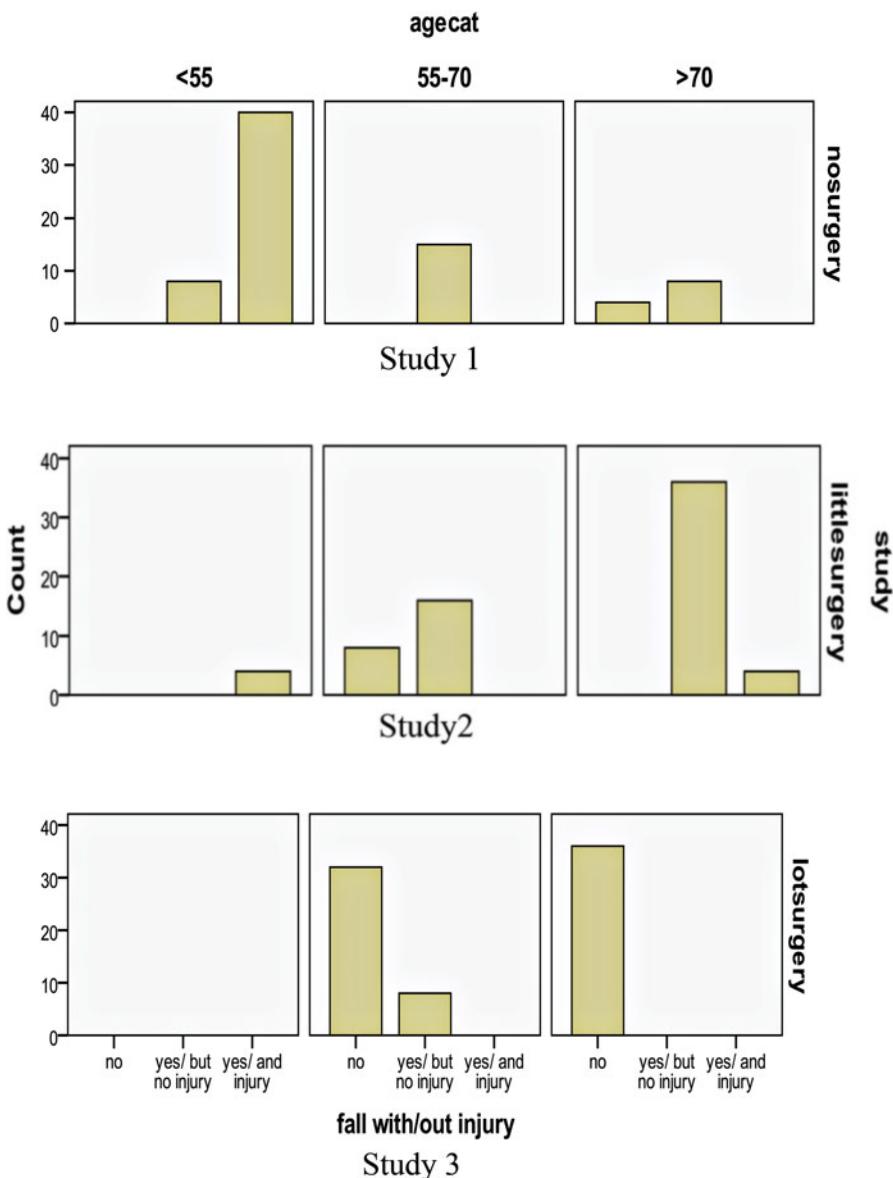
$$y = a + 3b_1 + b_2x_2$$

Each  $x_1$ -value has its own regression line, all lines constitute a regression plane: this is the best fit plane (with the shortest distance to values in space). All of the regression lines have identical regression coefficients, the  $b_2$  - values, but the intercepts, given by  $a$ ,  $a + b_1$ ,  $a + 2b_1$ ,  $a + 3b_1$ , change continually. This is called multiple linear regression, but can also be called fixed effect intercept modeling.

In practice, the regression plane has a lot of uncertainty, caused by the standard deviations of  $x_1$  and  $x_2$ , and, in addition, some unexplained uncertainty, otherwise called residual uncertainty. With null hypothesis testing we, usually, test whether the amount of uncertainty due to the variables  $x_1$  and  $x_2$  is considerably larger than that due to residual uncertainty. This is called a fixed effect intercept testing. Sometimes a better sensitivity of testing is obtained by a random effect intercept modeling or random intercept modeling. Here the standard deviation of for example  $x_1$  is assumed to be unexpected, not something we could predict, and it is, therefore, added to the residual uncertainty of the model. With truly unexpected effects, often, indeed, a better datafit is obtained, and therefore better p-values of testing.

## 13.2 Example, Meta-analysis of Three Studies

The three rows of graphs give underneath show three studies with histograms of patients by ageclasses. The three studies assess the effect of ageclass on fall out of bed while in hospital



The above graphs of bars of counts of falloutofbed shows, that the three studies were very heterogeneous. Particularly, some cells were empty or nearly empty. For

example, in the bottom row study no patients under 55 were present, and in the top row study patients under 55 without falloutofbed were never observed, and neither were patients over 70 with falloutofbed and injury. The type of patients, from no surgery to a lot of surgery, might be responsible for the heterogeneity, because surgery, particularly, in older patients may be accompanied by increased risks of fall out of bed. Therefore, it was decided to perform a random intercept analysis for categorical outcome and predictor variables, with the falloutofbed as outcome, and age category and study as respectively fixed effect and random effect predictor variables.

Categories are very common in medical research. Examples include age classes, income classes, education levels, drug dosages, diagnosis groups, disease severities, etc. Statistics has generally difficulty to assess categories, and traditional models require either binary or continuous variables. If in the outcome, categories can be assessed with multinomial regression (SPSS for starters and 2nd levelers second edition, Chap. 44, Multinomial regression for outcome categories, Springer Heidelberg Germany, 2016, from the same authors). If as predictors, they can be assessed with linear regression for categorical predictors (SPSS for starters and 2nd levelers second edition, Chap. 8, Linear regression with categorical predictors, Springer Heidelberg Germany, 2016, from the same authors). However, with multiple categories or with categories both in the outcome and as predictors, random intercept models may provide better sensitivity of testing. The latter models assume that for each predictor category or combination of categories  $x_1, x_2, \dots$  slightly different  $a$ -values can be computed with a better fit for the outcome category  $y$  than a single  $a$ -value.

$$y = a + b_1x_1 + b_2x_2 + \dots$$

We should add that, instead of the above linear equation, even better results are often obtained with log-transformed outcome variables ( $\log$  = natural logarithm).

$$\log y = a + b_1x_1 + b_2x_2 + \dots$$

Are in a study of exposure and outcome categories the exposure categories significant predictors of the outcome categories, and does a random intercept provide better test-statistics than does a fixed effects analysis?

The outcome data of the three above hospital trials were very heterogeneous as shown in the above 3 graphs, and the heterogeneity was ascribed to type of patients, those receiving (1) no surgery, (2) little surgery, (3) a lot of surgery. They were, therefore, included in a random intercept model as predictor category. Also three patient age classes (1 young, 2 middle, 3 old) were assumed to be predictor categories. Are the predictor categories significant determinants of the outcome categories risk of falling out of bed (1) no, (2) with or (3) without injury, and does a random intercept provide better statistics than a fixed effect categorical analysis model?

The data file is in extras.[springer.com](http://springer.com) and is entitled “randomintercepts”. SPSS version 20 and up can be used for analysis. First, we will perform a fixed intercept model.

The module Mixed Models consists of two statistical models:

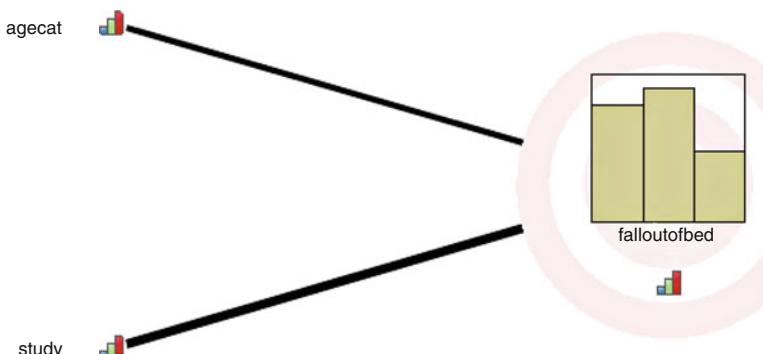
Linear,  
Generalized Linear.

For analysis the statistical model Generalized Linear Mixed Models is required. First, we will perform a fixed effects model analysis, then a random effect model.

### Command

click Analyze....Mixed Models....Generalized Linear Mixed Models....click Data Structure....click "patient\_id" and drag to Subjects on the Canvas....click Fields and Effects....click Target....Target: select "fall with/out injury"....click Fixed Effects ....click "agecat"and "study" and drag to Effect Builder:....mark Include intercept....click Run.

The underneath results show, that both the various regression coefficients and the overall correlation coefficients between the predictors and the outcome are, generally, statistically significant.

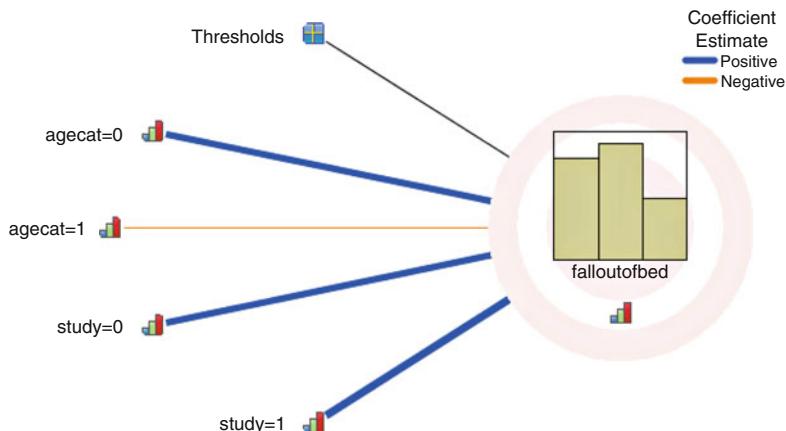


Source	F	df1	df2	Sig.
Corrected Model ▼	9,496	4	10	,002
agecat	6,914	2	10	,013
study	9,924	2	10	,004

Probability distribution: Multinomial  
Link function: Cumulative logit

In the mixed model output sheets an interactive graph is observed with predictors shown as lines with thicknesses corresponding to their predictive power and the outcome in the form of a histogram. There is no best fit Gaussian curve in the above histogram, because a nongaussian multinomial distribution has been chosen (Chap. 44, SPSS for starters and 2nd levelers, Springer Heidelberg Germany, from the same authors). Overall, both the predictors "agecat and study" are statistically significant, respectively at p-values of 0.013 and 0.004.

The underneath graph shows the effects of the single categorical variables. Again many variables were significant.



Model Term	Coefficient ►	Sig.
Threshold for falloutofbed = 0	2,133	,027
agecat=1	7,203	,000
agecat=0	5,243	,005
agecat=1	-0,016	,986
agecat=2	0,000 <sup>a</sup>	
study=0	3,627	,008
study=1	4,260	,002
study=2	0,000 <sup>a</sup>	

Probability distribution: Multinomial  
Link function: Cumulative logit

<sup>a</sup>This coefficient is set to zero because it is redundant.

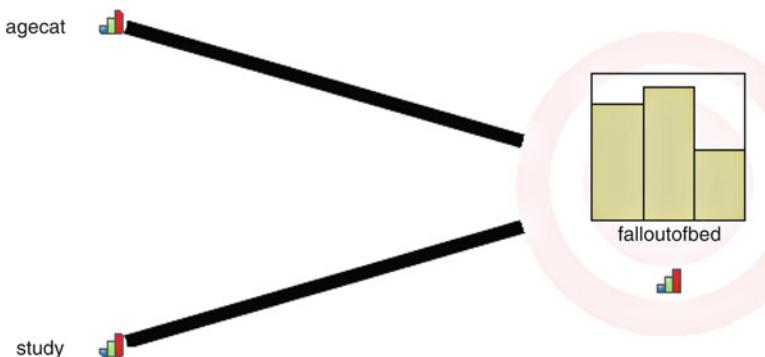
Subsequently, a random intercept analysis will be performed.

### Command

Analyze....Mixed Models....Generalized Linear Mixed Models....click Data Structure....click "patient\_id" and drag to Subjects on the Canvas....click Fields and Effects....click Target: select "fall with/out injury"....click Fixed Effects....click "agecat" and "study" and drag to Effect Builder:....mark Include intercept....click Random Effects....click Add Block....mark Include intercept....Subject combination: select patient\_id....click OK....click Model Options....click Save Fields....mark PredictedValue....mark PredictedProbability....click Save ....click Run.

The underneath results show the test-statistics of the random intercept model. The random intercept model shows better statistics:

p = 0.000 and 0.013 overall for age,  
 p = 0.000 and 0.004 overall for study,  
 p = 0.000 and 0.005 regression coefficients for age class 0 versus 2,  
 p = 0.814 and 0.998 for age class 1 versus 2,  
 p = 0.000 and 0.008 for study 0 versus 2, and  
 p = 0.000 and 0.0002 for study 1 versus 2.

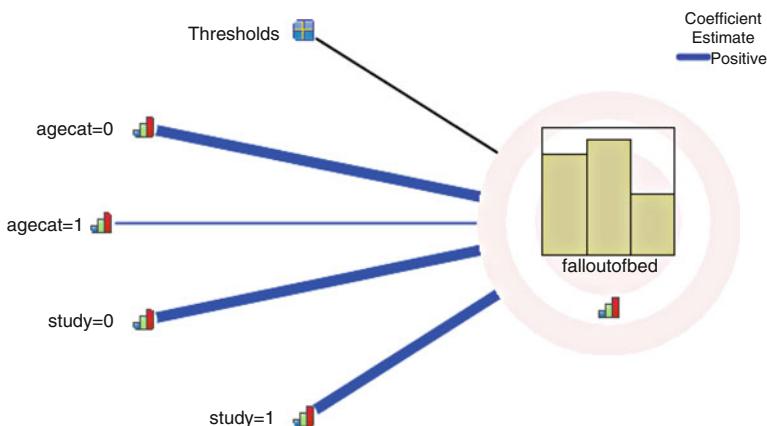


Source	F	df1	df2	Sig.
Corrected Model ▼	31,583	4	213	,000
agecat	21,803	2	213	,000
study	30,181	2	213	,000

Probability distribution: Multinomial  
 Link function: Cumulative logit

If no adjustment for multiple testing a Bonferroni p-value of  $0.05 \times 2/[k(k - 1)]$ , instead of 0.050 was calculated, then the rejecting p-value with 6 p-values would turn out to be 0.0033. The overall test for predicting age and the study, and the regression coefficient for age class 0 versus 2, and for study 0 versus 2 would no

longer be statistically significant. In contrast, with the random effect model the adjustment would not affect any of these statistical significance tests.



Model Term	Coefficient ►	Sig.
<b>Threshold for falloutofbed=</b>	0	2,079 ,000
	1	5,447 ,000
<b>agecat=0</b>	3,872	,000
<b>agecat=1</b>	0,090	,814
<b>agecat=2</b>	0,000 <sup>a</sup>	
<b>study=0</b>	3,208	,000
<b>study=1</b>	3,559	,000
<b>study=2</b>	0,000 <sup>a</sup>	

Probability distribution: Multinomial

Link function: Cumulative logit

<sup>a</sup>This coefficient is set to zero because it is redundant.

Like automatic linear regression (see SPSS for starters and 2nd levelers second edition, Chap. 7, Automatic linear regression, Springer Heidelberg Germany, 2016, from the same authors), and other generalized mixed linear models (see SPSS for starters and 2nd levelers second edition, Chap. 12, Repeated measures mixed-modeling, Springer Heidelberg Germany, 2016, from the same authors), random intercept models include the possibility to make XML files from the analysis, that can, subsequently, be used for making predictions about the chance of falling out of bed in future patients. However, SPSS uses here slightly different software called winRAR ZIP files that are “shareware”. This means that you pay a small fee and be registered if you wish to use it. Note that winRAR ZIP files have an archive file format consistent of compressed data used by Microsoft since 2006 for the purpose of filing XML (eXtended Markup Language) files. They are only employable for a limited period of time like e.g. 40 days.

### 13.3 Conclusion

Generalized linear mixed models are suitable for analyzing data with multiple categorical variables. Random intercept versions of these models provide better sensitivity of testing than fixed intercept models, and are adequate for testing meta-analyses of studies with both outcome and predictor categories. The current chapter uses an example of a meta-analysis of just three studies. We should add that random intercepts meta-analysis has additional benefits. Particularly with meta-analyses of very large numbers of studies, it can help selecting the worthwhile studies and skipping the rest. With jackknife resampling of the studies, the recombinations of studies with the best predictive statistics can be easily identified. The procedure is called capturing gravity (Gee, The concept of gravity in meta-analysis, Counseling, Psychotherapy and Health 2005; 1: 52–75). This methodology is, particularly, popular in social science research, where multiple studies are more common than in clinical research. But, some bioclinical meta-analyses have been published (Field et al, Spatial species-richness across scales a meta-analysis, J Biogeography 2008; doi 10.1111, and Lukacik et al, A meta-analysis of effects of oral zinc in the treatment of diarrhoea, Pediatrics 2008; 121: 102).

## Reference

More information on statistical methods for analyzing data with categories is, e.g., in the Chaps. 8, 39, and 44, SPSS for starters and 2nd levelers second edition, Springer Heidelberg Germany, 2016, from the same authors.

# **Chapter 14**

## **Probit Meta-regression**

### **Dose Response Meta-analyses**

**Abstract** If your predictor is multiple pharmacological treatment dosages, then probit regression may be more convenient than (multinomial) logistic regression, because your results will be reported in the form of response rates instead of odds ratios.

As an example, in a dose response meta-analysis of 14 mosquito studies with different dosages of chemical (chem) and nonchemical (nonchem) repellents the numbers of mosquitos gone after administration were assessed and meta-analyzed. The probit regression model provided adequate power for comparing the effects of different dosages.

#### **14.1 Introduction**

Probit regression is for estimating the effect of predictors on yes/no outcomes. If your predictor is multiple pharmacological treatment dosages, then probit regression may be more convenient than logistic regression, because your results will be reported in the form of response rates instead of odds ratios. The dependent variable of the two methods log odds (otherwise called logit) and log prob. (otherwise called probit) are closely related to one another. It can be shown that the log odds of responding  $\approx (\pi/\sqrt{3}) \times \log$  probability of responding (Machine learning in medicine part three, Chap. 7, Probit regression, pp. 63–68, 2013, Springer Heidelberg Germany, from the same authors). This chapter will assess, whether probit regression is able to test whether meta-data of studies with different predictor levels, like dose-response levels, can be used for comparing response rates.

#### **14.2 Example**

As an example, in a dose response meta-analysis of 14 mosquito studies with different dosages of chemical (chem) and nonchemical (nonchem) repellents the numbers of mosquitos gone after administration were assessed and meta-analyzed.

study number	mosquitos gone	study size mosquitos	repellent nonchem	repellent chem
1	1000	18,000	1	,02
2	1000	18,500	1	,03
3	3500	19,500	1	,03
4	4500	18,000	1	,04
5	9500	16,500	1	,07
6	17,000	22,500	1	,09
7	20,500	24,000	1	,10
8	500	22,500	2	,02
9	1500	18,500	2	,03
10	1000	19,000	2	,03
11	5000	20,000	2	,04
12	10,000	22,000	2	,07
13	8000	16,500	2	,09
14	13,500	18,500	2	,10

For analysis the statistical model Probit Regression in the module Regression of SPSS statistical software is required. Start by entering the above data file in the Data Viewer Screen of the software program.

### Command

Analyze....Regression....Probit Regression....Response Frequency: enter "mosquitos gone"....Total Observed: enter "n mosquitos"....Covariate(s): enter "chemical"....Transform: select "natural log"....click OK.

### Chi-Square Tests

		Chi-Square	df <sup>a</sup>	Sig.
PROBIT	Pearson Goodness-of-Fit Test	7706,816	12	,000 <sup>b</sup>

- a. Statistics based on individual cases differ from statistics based on aggregated cases.
- b. Since the significance level is less than ,150, a heterogeneity factor is used in the calculation of confidence limits.

In the output sheets the above table shows, that the goodness of fit tests of the data is statistically significant, and, thus, that the data do not fit the probit model very well. However, SPSS is going to produce a heterogeneity correction factor, and, so, we can proceed. The underneath tables in the output sheets show, that chemical dilution levels are a very significant predictor of the proportions of mosquitos gone.

### Parameter Estimates

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
PROBIT <sup>a</sup>	chemical (dilution)	,006	286,098	,000	1,638	1,660
	Intercept	,017	267,094	,000	4,472	4,506

a. PROBIT model: PROBIT( $p$ ) = Intercept + BX (Covariates X are transformed using the base 2.718 logarithm.)

### Cell Counts and Residuals

	Number	chemical (dilution)	Number of Subjects	Observed Responses	Expected Responses	Residual	Probability
PROBIT	1	-3,912	18000	1000	448,194	551,806	,025
	2	-3,624	18500	1000	1266,672	-266,672	,068
	3	-3,401	19500	3500	2564,259	935,741	,132
	4	-3,124	18000	4500	4574,575	-74,575	,254
	5	-2,708	16500	9500	8405,866	1094,134	,509
	6	-2,430	22500	17000	15410,676	1589,324	,685
	7	-2,303	24000	20500	18134,992	2365,008	,756
	8	-3,912	22500	500	560,243	-60,243	,025
	9	-3,624	18500	1500	1266,672	233,328	,068
	10	-3,401	19000	1000	2498,508	-1498,508	,132
	11	-3,124	20000	5000	5082,861	-82,861	,254
	12	-2,708	22000	10000	11207,821	-1207,821	,509
	13	-2,430	16500	8000	11301,162	-3301,162	,685
	14	-2,303	18500	13500	13979,056	-479,056	,756

The above table shows, that, according to chi-square tests, the differences between observed and expected proportions of mosquitos gone is several times statistically significant at  $p = 0.025$  and  $0.068$ , with  $p = 0.20$  as threshold for significance.

It does, therefore, make sense to make some inferences using the underneath confidence limits table.

Confidence Limits

Probability	95% Confidence Limits for chemical (dilution)			95% Confidence Limits for log(chemical (dilution)) <sup>b</sup>		
	Estimate	Lower Bound	Upper Bound	Estimate	Lower Bound	Upper Bound
.010	,016	,012	,020	-4,133	-4,453	-3,911
,020	,019	,014	,023	-3,968	-4,250	-3,770
,030	,021	,016	,025	-3,863	-4,122	-3,680
,040	,023	,018	,027	-3,784	-4,026	-3,612
,050	,024	,019	,029	-3,720	-3,949	-3,557
,060	,026	,021	,030	-3,665	-3,882	-3,509
,070	,027	,022	,031	-3,617	-3,825	-3,468
,080	,028	,023	,032	-3,574	-3,773	-3,430
,090	,029	,024	,034	-3,535	-3,726	-3,396
,100	,030	,025	,035	-3,500	-3,683	-3,365
,150	,035	,030	,039	-3,351	-3,506	-3,232
,200	,039	,034	,044	-3,233	-3,368	-3,125
,250	,044	,039	,048	-3,131	-3,252	-3,031
,300	,048	,043	,053	-3,040	-3,150	-2,943
,350	,052	,047	,057	-2,956	-3,059	-2,860
,400	,056	,051	,062	-2,876	-2,974	-2,778
,450	,061	,055	,067	-2,799	-2,895	-2,697
,500	,066	,060	,073	-2,722	-2,819	-2,614
,550	,071	,064	,080	-2,646	-2,745	-2,529
,600	,077	,069	,087	-2,569	-2,672	-2,442
,650	,083	,074	,095	-2,489	-2,598	-2,349
,700	,090	,080	,105	-2,404	-2,522	-2,251
,750	,099	,087	,117	-2,313	-2,441	-2,143
,800	,109	,095	,132	-2,212	-2,351	-2,022
,850	,123	,106	,153	-2,094	-2,248	-1,879
,900	,143	,120	,183	-1,945	-2,120	-1,699
,910	,148	,124	,191	-1,909	-2,089	-1,655
,920	,154	,128	,200	-1,870	-2,055	-1,608
,930	,161	,133	,211	-1,827	-2,018	-1,556
,940	,169	,138	,224	-1,780	-1,977	-1,497
,950	,178	,145	,239	-1,725	-1,931	-1,430
,960	,190	,153	,259	-1,661	-1,876	-1,352
,970	,206	,164	,285	-1,582	-1,809	-1,255
,980	,228	,179	,324	-1,477	-1,719	-1,126
,990	,269	,206	,397	-1,312	-1,579	-923

a. A heterogeneity factor is used.

b. Logarithm base = 2.718.

E.g., one might conclude, that a 0.143 dilution of the chemical repellent causes 0.900 (=90%) of the mosquitos to have gone. And a 0.066 dilution would mean, that 0.500 (=50%) of the mosquitos disappeared.

Just like linear and logistic regression models, probit regression can also be applied with multiple predictors. We will add as second predictor to the above example: the nonchemical repellents ultrasound (= 1) and burning candles (= 2). See table underneath.

study number	mosquitos gone	study size	repellent nonchem	repellent chem
1	1000	18,000	1	,02
2	1000	18,500	1	,03
3	3500	19,500	1	,03
4	4500	18,000	1	,04
5	9500	16,500	1	,07
6	17,000	22,500	1	,09
7	20,500	24,000	1	,10
8	500	22,500	2	,02
9	1500	18,500	2	,03
10	1000	19,000	2	,03
11	5000	20,000	2	,04
12	10,000	22,000	2	,07
13	8000	16,500	2	,09
14	13,500	18,500	2	,10

### Command

Analyze....Regression....Probit Regression....Response Frequency: enter "mosquitos gone"....Total Observed: enter "n mosquitos"....Covariate(s): enter "chemical, nonchemical"....Transform: select "natural log"....click OK.

### Chi-Square Tests

		Chi-Square	df <sup>a</sup>	Sig.
PROBIT	Pearson Goodness-of-Fit Test	3863,489	11	,000 <sup>b</sup>

a. Statistics based on individual cases differ from statistics based on aggregated cases.

b. Since the significance level is less than ,150, a heterogeneity factor is used in the calculation of confidence limits.

Again, the goodness of fit is not what it should be, but SPSS adds a correction factor for heterogeneity. The underneath table shows the regression coefficients for the multiple model. The nonchemical repellents have significantly different effects on the outcome.

## Parameter Estimates

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
PROBIT <sup>a</sup> chemical (dilution)	1,654	,006	284,386	,000	1,643	1,665
Intercept <sup>b</sup> ultrasound	4,678	,017	269,650	,000	4,661	4,696
burning candles	4,321	,017	253,076	,000	4,304	4,338

a. PROBIT model: PROBIT(p) = Intercept + BX (Covariates X are transformed using the base 2.718 logarithm.)

b. Corresponds to the grouping variable repellentnonchemical.

## Cell Counts and Residuals

	Number	repellentnonchemical	chemical (dilution)	Number of Subjects	Observed Responses	Expected Responses	Residual	Probability
PROBIT	1	1	-3,912	18000	1000	658,233	341,767	,037
	2	1	-3,624	18500	1000	1740,139	-740,139	,094
	3	1	-3,401	19500	3500	3350,108	149,892	,172
	4	1	-3,124	18000	4500	5630,750	-1130,750	,313
	5	1	-2,708	16500	9500	9553,811	-53,811	,579
	6	1	-2,430	22500	17000	16760,668	239,332	,745
	7	1	-2,303	24000	20500	19388,521	1111,479	,808
	8	2	-3,912	22500	500	355,534	144,466	,016
	9	2	-3,624	18500	1500	871,485	628,515	,047
	10	2	-3,401	19000	1000	1824,614	-824,614	,096
	11	2	-3,124	20000	5000	3979,458	1020,542	,199
	12	2	-2,708	22000	10000	9618,701	381,299	,437
	13	2	-2,430	16500	8000	10202,854	-2202,854	,618
	14	2	-2,303	18500	13500	12873,848	626,152	,696

In the above Cell Counts table, also given in the output, it is shown, that, according to the chi-square tests, the differences of observed and expected proportions of mosquitos gone were statistically significant several times at  $p = 0.016$ ,  $0.037$ ,  $0.047$ ,  $0.096$ .

The next page tables give interesting results. E.g., a 0.128 dilution of the chemical repellent causes 0.900 (=90%) of the mosquitos to have gone in the ultrasound tests. And 0.059 dilution would mean that 0.500 (=50%) of the mosquitos disappeared. The results of burning candles were less impressive. 0.159 Dilution caused 90% of the mosquitos to disappear, a 0.073 dilution did 50% to do so.

## Confidence Limits

nonchemical	Probability	95% Confidence Limits for chemical (dilution)			95% Confidence Limits for log(chemical (dilution)) <sup>b</sup>		
		Estimate	Lower Bound	Upper Bound	Estimate	Lower Bound	Upper Bound
ultrasound	,010	,014	,011	,018	-4,235	-4,486	-4,042
	,020	,017	,014	,020	-4,070	-4,296	-3,895
	,030	,019	,015	,022	-3,966	-4,176	-3,801
	,040	,021	,017	,024	-3,887	-4,086	-3,731
	,050	,022	,018	,025	-3,823	-4,013	-3,673
	,060	,023	,019	,027	-3,769	-3,951	-3,624
	,070	,024	,020	,028	-3,721	-3,896	-3,581
	,080	,025	,021	,029	-3,678	-3,848	-3,542
	,090	,026	,022	,030	-3,639	-3,804	-3,506
	,100	,027	,023	,031	-3,603	-3,763	-3,473
	,150	,032	,027	,036	-3,455	-3,597	-3,337
	,200	,036	,031	,040	-3,337	-3,467	-3,227
	,250	,039	,035	,044	-3,236	-3,356	-3,131
	,300	,043	,038	,048	-3,146	-3,258	-3,043
	,350	,047	,042	,052	-3,062	-3,169	-2,961
	,400	,051	,046	,056	-2,982	-3,085	-2,882
	,450	,055	,049	,061	-2,905	-3,006	-2,803
	,500	,059	,053	,066	-2,829	-2,929	-2,725
	,550	,064	,058	,071	-2,753	-2,853	-2,646
	,600	,069	,062	,077	-2,675	-2,777	-2,564
	,650	,075	,067	,084	-2,596	-2,700	-2,478
	,700	,081	,073	,092	-2,512	-2,620	-2,387
	,750	,089	,079	,102	-2,421	-2,534	-2,287
	,800	,098	,087	,114	-2,320	-2,440	-2,174
	,850	,111	,097	,130	-2,202	-2,332	-2,042
	,900	,128	,111	,153	-2,054	-2,197	-1,874
	,910	,133	,115	,160	-2,018	-2,165	-1,833
	,920	,138	,119	,167	-1,979	-2,129	-1,789
	,930	,144	,124	,175	-1,936	-2,091	-1,740
	,940	,151	,129	,185	-1,889	-2,048	-1,686
	,950	,160	,135	,197	-1,834	-1,999	-1,623
	,960	,170	,143	,212	-1,770	-1,942	-1,550
	,970	,184	,154	,232	-1,691	-1,871	-1,459
	,980	,205	,169	,262	-1,587	-1,778	-1,339
	,990	,241	,196	,317	-1,422	-1,632	-1,149

## Confidence Limits

nonchemical	Probability	95% Confidence Limits for chemical (dilution)			95% Confidence Limits for $\log(\text{chemical dilution})^b$		
		Estimate	Lower Bound	Upper Bound	Estimate	Lower Bound	Upper Bound
burning candles	,010	,018	,014	,021	-4,019	-4,247	-3,841
	,020	,021	,017	,025	-3,854	-4,058	-3,693
	,030	,024	,019	,027	-3,750	-3,939	-3,599
	,040	,025	,021	,029	-3,671	-3,850	-3,528
	,050	,027	,023	,031	-3,607	-3,777	-3,469
	,060	,029	,024	,033	-3,553	-3,716	-3,420
	,070	,030	,026	,034	-3,505	-3,662	-3,376
	,080	,031	,027	,036	-3,462	-3,614	-3,336
	,090	,033	,028	,037	-3,423	-3,571	-3,300
	,100	,034	,029	,038	-3,387	-3,531	-3,267
	,150	,039	,034	,044	-3,239	-3,367	-3,128
	,200	,044	,039	,049	-3,121	-3,240	-3,015
	,250	,049	,044	,054	-3,020	-3,132	-2,916
	,300	,053	,048	,059	-2,930	-3,037	-2,826
	,350	,058	,052	,065	-2,845	-2,950	-2,741
	,400	,063	,057	,070	-2,766	-2,869	-2,658
	,450	,068	,061	,076	-2,688	-2,793	-2,578
	,500	,073	,066	,082	-2,613	-2,718	-2,497
	,550	,079	,071	,089	-2,537	-2,644	-2,415
	,600	,085	,076	,097	-2,459	-2,571	-2,331
	,650	,093	,082	,106	-2,380	-2,495	-2,244
	,700	,101	,089	,116	-2,295	-2,417	-2,151
	,750	,110	,097	,129	-2,205	-2,333	-2,049
	,800	,122	,106	,144	-2,104	-2,240	-1,936
	,850	,137	,119	,165	-1,986	-2,133	-1,802
	,900	,159	,136	,195	-1,838	-1,999	-1,633
	,910	,165	,140	,203	-1,802	-1,966	-1,592
	,920	,172	,145	,213	-1,763	-1,932	-1,548
	,930	,179	,151	,223	-1,720	-1,893	-1,499
	,940	,188	,157	,236	-1,672	-1,850	-1,444
	,950	,198	,165	,251	-1,618	-1,802	-1,381
	,960	,211	,175	,270	-1,554	-1,745	-1,308
	,970	,229	,187	,296	-1,475	-1,675	-1,217
	,980	,254	,206	,334	-1,371	-1,582	-1,096
	,990	,299	,238	,404	-1,206	-1,436	-906

### 14.3 Conclusion

The meta-analysis of dose response studies is not straightforward, because dose response curves are hard to meta-analyze. This issue has raised already discussions in the past. For example, Berlin et al. (Epidemiology 1993; 4: 218–228) stated, that, given the between-study heterogeneities, dose response meta-analyses are much harder to meta-analyze than single dose-response studies are. They recommended relative risks, rate ratios or odds ratios as outcome measures, rather than the traditional within-study dose response slopes. This was supported by the biostatistician Dunouchel from New York(Stat Med 1995; 14: 679–685), showing, that the appropriate model for dose response meta-analyses may not be linear regression modeling, and, that, given the heterogeneity between studies, dose response meta-analyse are arithmetically much harder than single dose dose-response studies. The current version of the Cochrane handbook of meta-analyses stated in its online general methods reviews ([handbook.cochrane.org](http://handbook.cochrane.org)), that claims about differences based on between study differences require caution. Particularly so, if they do not exist in the regression models, that are, notoriously, low power. The probit regression, as applied in the current chapter, provides much more power, and, so, it is a welcome alternative to the purpose.

In conclusion, probit regression can be adequately used for meta-analyzing the effects of different dosages of chemical and nonchemical repellents on death rates different populations of mosquitos. If your predictor is multiple pharmacological treatment dosages, then probit regression may be more convenient than logistic regression, because results will be reported in the form of response rates instead of odds ratios. This chapter shows that probit regression is adequate for comparing different response rates of studies using different dosages of mosquito repellents.

## Reference

More background, theoretical and mathematical information of probit regression is given in Machine learning in medicine part three, Chap.7, Probit regression, pp 63–68, 2013, Springer Heidelberg Germany (from the same authors).

# **Chapter 15**

## **Meta-analysis with General Loglinear Models**

### **Regression with Weighted Least Squares and General Loglinear Models**

**Abstract** General loglinear modeling can be used for assessing the effect of discrete predictors if such predictors interact with one another. In an example of 12 population studies of patients with different drinking patterns the effect of the drinking patterns on the risk of infarction were assessed. The methodology identified subgroups with significantly larger or smaller risks of infarction.

#### **15.1 Introduction**

Data files that assess the effect of discrete predictors on frequency counts of morbidities/mortalities can be assessed with multiple linear regression. However, the results do not mean too much, if the predictors interact with one another. In that case they can be cross-classified in tables of multiple cells using general loglinear modeling. Can general loglinear modeling identify subgroups with significantly larger incident risks than other subgroups.

#### **15.2 Example, Weighted Multiple Linear Regression**

As an example, in 12 studies of populations at risk of infarction with little or lot of soft drink consumption, and consumptions of wine and other alcoholic beverages the incident risk of infarction equaled  $240/930 = 24.2\%$ , in those with lots of soft drinks, no wine, and no alcohol otherwise it did  $285/1043 = 27.3\%$ .

soft drink (1 = little)	wine (0 = no)	alc beverages (0 = no)	infarcts number	population number
1,00	1,00	1,00	240	993
1,00	1,00	,00	237	998
2,00	1,00	1,00	236	1016
2,00	1,00	,00	236	1011
3,00	1,00	1,00	221	1004
3,00	1,00	,00	221	1003
1,00	,00	1,00	270	939
1,00	,00	,00	269	940
2,00	,00	1,00	274	979
2,00	,00	,00	273	966
3,00	,00	1,00	284	1041
3,00	,00	,00	285	1043

We wish to identify the populations with particular high risks. Start by opening the data file in SPSS. For analysis the statistical model Linear in the module Regression is required. Start by entering the data in the Data View Screen of the software program.

### Command

Analyze....Linear Regression....Dependent: infarcts....Independent(s): sof drink, wine, other alc (alcoholic) beverages....WLS Weight: population....click OK.

Coefficients<sup>a,b</sup>

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
	B	Std. Error	Beta			
1 (Constant)	277,397	,201		1381,647	,000	
soft drink	-,657	,080	-,023	-8,213	,000	
wine	-44,749	,131	-,953	-342,739	,000	
other alc beverages	,569	,130	,012	4,364	,000	

a. Dependent Variable: infarcts

b. Weighted Least Squares Regression - Weighted by population

The above tables show that the three discrete predictors soft drink, wine, and other “alc beverages” are very strong independent predictors of infarcts, and they are adjusted for population size. However, the regression coefficient of “other alcoholic beverages” is positive, indicating that the more consumption the higher the risk will be, while the regression coefficients of soft drinks and wine are negative, indicating the reverse to be true. We will now add the underneath interaction variables to the data:

interaction variable 1 = wine \* other alc beverages

interaction variable 2 = soft drink \* wine

interaction variable 3 = soft drink \* other alc beverages

### Command

The same commands are given as those given above. However, the interaction variables were added as additional predictors.

**ANOVA<sup>b,c</sup>**

Model	Sum of Squares	df	Mean Square	F	Sig.
1      Regression	5,941E9	6	9,902E8	19757,821	,000 <sup>a</sup>
Residual	5,976E8	11924	50118,453		
Total	6,539E9	11930			

a. Predictors: (Constant), soft\*alc, wine, soft drink, soft\*wine, wine\*alc, other alc beverages

b. Dependent Variable: infarcts

c. Weighted Least Squares Regression - Weighted by population

**Coefficients<sup>a,b</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1      (Constant)	275,930	,272		1013,835	,000
soft drink	,118	,115	,004	1,026	,305
wine	-45,619	,224	-,972	-203,982	,000
other alc beverages	3,762	,407	,080	9,240	,000
wine*alc	,103	,269	,002	,385	,700
soft*wine	,487	,096	,024	5,088	,000
soft*alc	-1,674	,175	-,084	-9,561	,000

a. Dependent Variable: infarcts

b. Weighted Least Squares Regression - Weighted by population

The output sheets now show, that soft drink is no longer a significant predictor of infarcts, while several interactions were very significant. This leaves us with a pretty inconclusive analysis. Due to the interactions the meaning of the former discrete predictors have little meaning any more. See for more information of interaction modeling Chap. 30, Statistics applied to clinical studies 5th edition, 2012, Springer Heidelberg Germany, from the same authors.

### 15.3 Example, General Loglinear Modeling

The general loglinear model computes cell counts in cross-classification tables, and can be simultaneously analyzed after logarithmic transformation in the form of analysis of variance data. In this way an overall analysis of subgroup differences can be produced, and the significant differences can be identified. For analysis the statistical model General Loglinear Analysis in the module Loglinear is required.

#### Command

Analyze....Loglinear ....General Loglinear Analysis....Factor(s): enter softdrink, wine, other alc beverages....click "Data" in the upper textrow of your screen.... click Weigh Cases....mark Weight cases by....Frequency Variable: enter "infarcts"....click OK....return to General Loglinear Analysis....Cell structure: enter "population".... Options ....mark Estimates....click Continue....Distribution of Cell Counts: mark Poisson....click OK.

The underneath tables are in the output sheets

Parameter Estimates<sup>b,c</sup>

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	-1,513	,067	-22,496	,000	-1,645	-1,381
[softdrink = 1,00]	,095	,093	1,021	,307	-,088	,278
[softdrink = 2,00]	,053	,094	,569	,569	-,130	,237
[softdrink = 3,00]	0 <sup>a</sup>	.	.	.	.	.
[wine = ,00]	,215	,090	2,403	,016	,040	,391
[wine = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[alcbeverages = ,00]	,003	,095	,029	,977	-,184	,189
[alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 1,00] * [wine = ,00]	-,043	,126	-,345	,730	-,291	,204
[softdrink = 1,00] * [wine = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 2,00] * [wine = ,00]	-,026	,126	-,209	,834	-,274	,221
[softdrink = 2,00] * [wine = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [wine = ,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [wine = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 1,00] * [alcbeverages = ,00]	-,021	,132	-,161	,872	-,280	,237
[softdrink = 1,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 2,00] * [alcbeverages = ,00]	,003	,132	,024	,981	-,256	,262
[softdrink = 2,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [alcbeverages = ,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + softdrink + wine + alcbeverages + softdrink \* wine + softdrink \* alcbeverages + wine \* alcbeverages + softdrink \* wine \* alcbeverages

Parameter Estimates<sup>b,c</sup>

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
[wine = ,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[wine = 1,00] * [alcbeverages = ,00]	0 <sup>a</sup>	.	.	.	.	.
[wine = 1,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 1,00] * [wine = ,00] * [alcbeverages = ,00]	,016	,178	,089	,929	-,334	,366
[softdrink = 1,00] * [wine = ,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 1,00] * [wine = 1,00] * [alcbeverages = ,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 1,00] * [wine = 1,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 2,00] * [wine = ,00] * [alcbeverages = ,00]	,006	,178	,036	,971	-,343	,356
[softdrink = 2,00] * [wine = ,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 2,00] * [wine = 1,00] * [alcbeverages = ,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 2,00] * [wine = 1,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [wine = ,00] * [alcbeverages = ,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [wine = ,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [wine = 1,00] * [alcbeverages = ,00]	0 <sup>a</sup>	.	.	.	.	.
[softdrink = 3,00] * [wine = 1,00] * [alcbeverages = 1,00]	0 <sup>a</sup>	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + softdrink + wine + alcbeverages + softdrink \* wine + softdrink \* alcbeverages + wine \* alcbeverages + softdrink \* wine \* alcbeverages

The above pretty dull tables give some wonderful comparisons between populations drinking various dosages of soft drinks, wine, and alcoholic beverages.

The soft drink classes 1 and 2 are not significantly different from zero. These classes have, thus, no greater risk of infarction than class 3. However, the regression coefficient of no wine is greater than zero at  $p = 0.016$ . No wine drinkers have a significantly greater risk of infarction than the wine drinkers have. No “other alcoholic beverages” did not protect from infarction better than the consumption of it. The three predictors did not display any interaction effects.

## 15.4 Conclusion

Studies with data files that assess the effects of discrete predictors on frequency counts of morbidities/mortalities can be meta-analyzed by classification of the studies into multiple cells of studies with varying incident risks (like,e.g., the incident risk of infarction) using general loglinear modeling.

They can identify subgroup studies with significantly larger or smaller incident risks than other subgroups. Linear regression can also be used for the purpose. However, possible interactions between the predictors require that interaction variable are computed and included in the linear model. Significant interaction variables render the linear regression model pretty meaningless (see also Chap. 23 of this edition).

## Reference

More background, theoretical and mathematical information of loglinear models are given in SPSS for starters and 2nd levelers 2nd edition, Chaps 51, Logit loglinear models, Chap. 52, Hierarchical loglinear models, Springer Heidelberg Germany, 2016, from the same authors. Interaction effects are reviewed in the Chap. 23 of the current edition.

# **Chapter 16**

## **Meta-analysis with Variance Components**

### **Reducing Unexplained Variance, Increasing Accuracy**

**Abstract** Variance components analysis is able to assess the magnitude of unexpected subgroup effects, otherwise called random effect, as compared to that of the residual error of a study. In three studies each of five patients the effect of three treatments on hours of sleep was assessed. Next, in ten studies each of four patients the effect of four treatments on the hours of sleep was assessed. Finally, in two studies each of 20 patients the effect of two treatments on the numbers of episodes of paroxysmal atrial tachycardias was assessed. Variance components analysis reduced the amount of unexpected variance, and increased accuracy.

### **16.1 Introduction**

If we have reasons to believe that in a study certain patients due to co-morbidity, co-medication and other factors will respond differently from others, then the spread in the data is caused not only by residual effect, but also by the subgroup properties, otherwise called random effect. Variance components analysis is able to assess the magnitudes of the random effect as compared to that of the residual error of a study. Can a variance components analysis by including the random effect in the analysis reduce the unexplained variance in a study, and, thus, increase the accuracy of the analysis model as used.

### **16.2 Example 1**

The data from three studies each of five patients assessed for hours of sleep during different treatments were meta-analyzed with variance components models. When the levels of a factor like treatment modalities have been chosen at random like treatments in phase I-II studies, investigators are often interested in components of variance. A components variance model is appropriate for the purpose.

A large numbers of treatment modalities have to be assessed and three parallel group studies of five sleeping sessions each will be performed for the purpose. The hours of sleep is the main outcome. The study samples are from a single population and have the same population variance, independent of sample size.

treatment modality	hours of sleep	study number	patient number
1,00	7,40	1,00	1
2,00	6,80	1,00	2
3,00	7,50	1,00	3
4,00	7,20	1,00	4
5,00	7,90	1,00	5
1,00	7,60	2,00	6
2,00	7,10	2,00	7
3,00	7,70	2,00	8
4,00	7,40	2,00	9
5,00	8,10	2,00	10
1,00	7,50	3,00	11
2,00	7,20	3,00	12
3,00	7,70	3,00	13
4,00	7,30	3,00	14
5,00	7,90	3,00	15

The mathematical model is  $y = \text{mean} + \text{mean variance}_{\text{population}} (\text{otherwise called } s^2_{\text{population}}) + \text{residual error.}$

The mean variance<sub>population</sub> is independent of sample size, and is an adequate estimate of the spread in the data. A one way analysis of variance (ANOVA) is performed with treatment modalities as outcome.

### ANOVA

outcome1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1,477	4	,369	20,519	,000
Within Groups	,180	10	,018		
Total	1,657	14			

The means squares give the error between treatments (0,369), and the residual error, (0,018). However, with random effect the errors are assessed differently:

$$\text{residual error} = s_e^2 = 0.018.$$

The mean square treatment now consists of the residual error plus three time the population

$$\text{variance} = s_e^2 + 3 s_{\text{population}}^2 = 0.369.$$

This would mean that:

$$s_{\text{population}}^2 = \frac{0.369 - 0.018}{3} = 0.117$$

and

$$s_{\text{total}}^2 = s_{\text{population}}^2 + s_e^2 = 0.117 + 0.018 = 0.135$$

and

$$0.117 / 0.135 = 86.7\%.$$

Thus, 86.7% of the total variance is due to differences between treatments, and 13.3% is due to differences within the treatment groups.

## 16.3 Example 2

The data from ten studies of four patients assessed for hours of sleep during different treatments were meta-analyzed with variance components models. When the levels of a factor like treatment modalities have been chosen at random like the treatments in the current example, investigators are often interested in components of variance. A components variance model will be performed again.

A large numbers of treatment modalities have to be assessed and ten parallel group studies of four sleeping sessions each will be performed for the purpose. The hours of sleep is the main outcome. The study samples are from a single population, and have, like in Example 1, the same population variance, independent of sample size.

treatment modality	hours of sleep	study number	patient number
,00	6,00	1,00	1
1,00	5,10	1,00	2
2,00	4,10	1,00	3
3,00	4,30	1,00	4
,00	7,10	2,00	5
1,00	8,00	2,00	6
2,00	7,00	2,00	7
3,00	4,30	2,00	8
,00	8,10	3,00	9
1,00	3,80	3,00	10
2,00	2,80	3,00	11
3,00	6,20	3,00	12
,00	7,50	4,00	13
1,00	4,40	4,00	14

2,00	3,40	4,00	15
3,00	5,60	4,00	16
,00	6,40	5,00	17
1,00	5,20	5,00	18
2,00	4,20	5,00	19
3,00	6,20	5,00	20
,00	7,90	6,00	21
1,00	5,40	6,00	22
2,00	4,40	6,00	23
3,00	6,00	6,00	24
,00	6,80	7,00	25
1,00	4,30	7,00	26
2,00	3,30	7,00	27
3,00	5,30	7,00	28
,00	6,60	8,00	29
1,00	6,00	8,00	30
2,00	5,00	8,00	31
3,00	5,40	8,00	32
,00	7,30	9,00	33
1,00	3,70	9,00	34
2,00	2,70	9,00	35
3,00	5,40	9,00	36
,00	5,60	10,00	37
1,00	6,20	10,00	38
2,00	5,20	10,00	39
3,00	5,30	10,00	40

The mathematical model is again  $y = \text{mean} + \text{mean variance}_{\text{population}}$  (otherwise called  $s^2_{\text{population}}$ ) + residual error. The mean variance is independent of sample size, and is an adequate estimate of the spread in the data. A one way analysis of variance is performed treatment modalities as outcome. Start by entering the above data in the Data Viewer Screen of SPSS statistical software.

### ANOVA

outcome2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	37,875	3	12,625	11,267	,000
Within Groups	40,339	36	1,121		
Total	78,214	39			

The mean squares give the error between treatments (12.625), and the residual error, (1.121). However, with random effect the errors are assessed differently:

$$\text{residual error} = s_e^2 = 1.121.$$

The mean square treatment now consists of the residual error plus three time the population

$$\text{variance} = s_e^2 + 3 s_{\text{population}}^2 = 12.625.$$

This would mean that:

$$s_{\text{population}}^2 = \frac{12.625 - 1.121}{10} = \frac{11.504}{10} = 1.15,$$

and

$$s_{\text{total}}^2 = s_{\text{population}}^2 + s_e^2 = 1.15 + 1.12 = 2.27,$$

and

$$1.15/2.27 = 0.506 = 50.6\%.$$

Thus, 50.6% of the total variance is due to differences between treatments and 49.4% due to differences within the treatment groups.

## 16.4 Example 3

The above two examples made use of the prior assumption of a single population variance. In practice this prior assumption rarely warranted, and also the variances of different studies will not be equal. In these cases, we will have to make use of the individual patient data of the studies. Special procedures are needed, requiring matrix algebra beyond this text in order to obtain ratios of observed and expected mean squares. SPSS statistical software version 20 and up is used for computations in the underneath example. An example is given of two 20 patient parallel-group studies assessing the effects of two treatments on the numbers of episodes of paroxysmal atrial fibrillation.

Variables	PAT	treat	gender	study
52,00	,00	,00	,00	2,00
48,00	,00	,00	,00	2,00
43,00	,00	,00	,00	1,00
50,00	,00	,00	,00	2,00
43,00	,00	,00	,00	2,00
44,00	,00	,00	,00	1,00
46,00	,00	,00	,00	2,00
46,00	,00	,00	,00	2,00
43,00	,00	,00	,00	1,00
49,00	,00	,00	,00	2,00

PAT = episodes of paroxysmal atrial tachycardias

treat = treatment modality (0 = placebo treatment, 1 = active treatment)

gender = gender (0 = female)

study number = study 1 no presence, study 2 yes presence of coronary artery disease.

The first 10 of the 40 patient data of the treatment of paroxysmal tachycardia with numbers of episodes of PAT as outcome is given above. The entire data file is in "variancecomponents", and is available at extras.[springer.com](http://springer.com). We had reason to believe, that the presence of coronary artery disease would affect the outcome, and, therefore, used this variable as a random rather than fixed variable. SPSS statistical software was used for data analysis. Start by opening the data file in SPSS.

### Command

Analyze....General Linear Model....Variance Components....Dependent Variable: enter "paroxtachyc"....Fixed Factor(s): enter "treat, gender"....Random Factor(s): enter "corardisease"....Model: mark Custom....Model: enter "treat, gender, cad".... click Continue....click Options....mark ANOVA....mark Type III....mark Sums of squares....mark Expected mean squares....click Continue....click OK.

The output sheets are given underneath. The Variance Estimate table gives the magnitude of the Variance due to cad, and that due to residual error (unexplained variance, otherwise called Error). The ratio of the Var (cad)/[Var (Error) + Var (cad)] gives the proportion of variance in the data due to the random cad effect  $(5.844/(28.426 + 5.844) = 0.206 = 20.6\%)$ . This means that 79.4% instead of 100% of the error is now unexplained.

**Variance Estimates**

Component	Estimate
Var(cad)	5,844
Var(Error)	28,426

Dependent Variable: paroxtach  
 Method: ANOVA  
 (Type III Sum of Squares)

The underneath ANOVA table gives the sums of squares and mean squares of different effects. E.g. the mean square of cad = 139.469, and that of residual effect = 28.426.

**ANOVA**

Source	Type III Sum of Squares	df	Mean Square
Corrected Model	727,069	3	242,356
Intercept	57153,600	1	57153,600
treat	515,403	1	515,403
gender	,524	1	,524
cad	139,469	1	139,469
Error	1023,331	36	28,426
Total	58904,000	40	
Corrected Total	1750,400	39	

Dependent Variable: paroxtach

The underneath Expected Mean Squares table gives the results of a special procedure, whereby variances of best fit quadratic functions of the variables are minimized to obtain the best unbiased estimate of the variance components. A little mental arithmetic is now required.

### Expected Mean Squares

Source	Variance Component		
	Var(cad)	Var(Error)	Quadratic Term
Intercept	20,000	1,000	Intercept, treat, gender
treat	,000	1,000	treat
gender	,000	1,000	gender
cad	19,000	1,000	
Error	,000	1,000	

Dependent Variable: paroxtach  
 Expected Mean Squares are based on Type III Sums of Squares.

For each source, the expected mean square equals the sum of the coefficients in the cells times the variance components, plus a quadratic term involving effects in the Quadratic Term cell.

EMS (expected mean square) of cad (the random effect)

$$\begin{aligned} &= 19 \times \text{Variance (cad)} + \text{Variance (Error)} \\ &= 139.469 \end{aligned}$$

EMS of Error (the residual effect)

$$\begin{aligned} &= 0 + \text{Variance (Error)} \\ &= 28.426 \end{aligned}$$

EMS of cad – Variance (Error)

$$\begin{aligned} &= 19 \times \text{Variance (cad)} \\ &= 139.469 - 28.426 \\ &= 110.043 \end{aligned}$$

Variance (cad)

$$\begin{aligned} &= 110.043/19 \\ &= 5.844 \text{ (compare with the results of the above Variance Estimates table)} \end{aligned}$$

It can, thus, be concluded that around 20% of the uncertainty is in the meta-analysis is caused by the random effect.

## 16.5 Conclusion

If we have reasons to believe, that, in a meta-analysis, certain patients due to co-morbidity, co-medication and other factors will respond differently from others, then the spread in the data will be caused, not only by the residual effect, but also by

the subgroup/substudy property, otherwise called the random effect. Variance components analysis, by including the random effect in the analysis, reduces the unexplained variance of the analysis, and, thus, increases its accuracy.

## Reference

More background, theoretical and mathematical information of random effect models are given in the chapter 4 of the current edition, and in Machine learning in medicine part three, Chap. 9, Random effect, pp 81–94, 2013, Springer Heidelberg Germany, from the same authors.

# **Chapter 17**

## **Ensembled Correlation Coefficients**

### **Less Noise but More Risk of Overfit**

**Abstract** Ensemble learning refers to the simultaneous use of multiple learning algorithms for the purpose of a better predictive power. Using SPSS Modeler a 250 patient data file with 28 variables, mainly patients' gene expression levels, were analyzed with linear regression, generalized linear model, nearest neighbor clustering, support vector machines, decision trees, chi-squares models, and neural networks. The correlation coefficients of the three best models were used for computing an average score and its errors. This average score was a lot better than those of the separate algorithms.

### **17.1 Introduction**

The term ensemble learning has been used by many authors from the machine learning community since the early 1990s. It refers to the simultaneous use of multiple learning algorithms for the purpose of obtaining a better predictive potential. Why should a better predictive potential from multiple algorithms be obvious? One may argue that the aggregate of a set of models is less noisy than the result of a single model. However, what about overfit of the aggregate causing statistical tests to be insignificant and inadequate. Mathematically, however, it can be shown that if individual models do not overfit, there will be no room for overfit of the aggregate. In addition it can be argued that often different models as used have their own way of classifying data. In line with Occam, the English Franciscan monk from the thirteenth century, traditional statistics says: the simplest model provides the best power. However, in line with Epicurus from Athens 300 BC, the machine learning community started to doubt the traditional concept, and started to think differently: if multiple analytic models explain data, keep them all.

Regarding the argument that different models as used have their own way of classifying data, an example will be given. In the data example of the next paragraph 12 highly expressed genes were the predictors of the outcome, a drug efficacy score. SPSS Modeler (see also Chap 48, Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors) was used to select the best fit models and perform an ensembled analysis. The three best fit models were

- (1) CHAID (chi square interaction detection) (see also Chap 8, Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors)
- (2) SVM (support vector machines) (see also Chap 71, Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors)
- (3) Linear Regression.

Linear regression produces a predictor outcome classification according to a straight line. Support vector machines works differently. It classifies outcome data in clusters, that are separated by socalled difficult observations, i.e., the observations closest to the separation lines. CHAID classifies the outcome according to a decision tree. It is a more complex method, extending a prior hypothesis of correlation to that of causality, and this is accomplished by a network of variables tested versus one another with the help of standardized rather than unstandardized regression coefficients. The three methodologies selected by the computer, have, obviously, ways of classifying data that are far from similar. This is furthermore supported by the example given underneath, that will show that they produced largely different numbers of analysis steps, here called fields, and different outcome values and error measures.

## 17.2 Ensemble Learning with SPSS Modeler

SPSS modeler is a work bench for automatic data mining (Chap. 61, Machine learning in medicine, Springer Heidelberg Germany, 2015, from the same authors) and modeling (see also the Chap. 64 and 65, Machine learning in medicine, Springer Heidelberg Germany, 2015, from the same authors). So far it is virtually unused in medicine, and mainly applied by econo-/sociometrics. Automatic modeling of continuous outcomes computes the ensembled result of a number of best fit models for a particular data set, and provides usually better sensitivity than the best fit separate models do. This chapter is to demonstrate whether it can be successfully applied for assessing the predictive potential of gene expression profiles on drug efficacy. The expression of a cluster of genes can be used as a functional unit to predict the efficacy of cytostatic treatment. Can ensembled modeling with three best fit statistical models provide better precision than the separate analysis with single statistical models does.

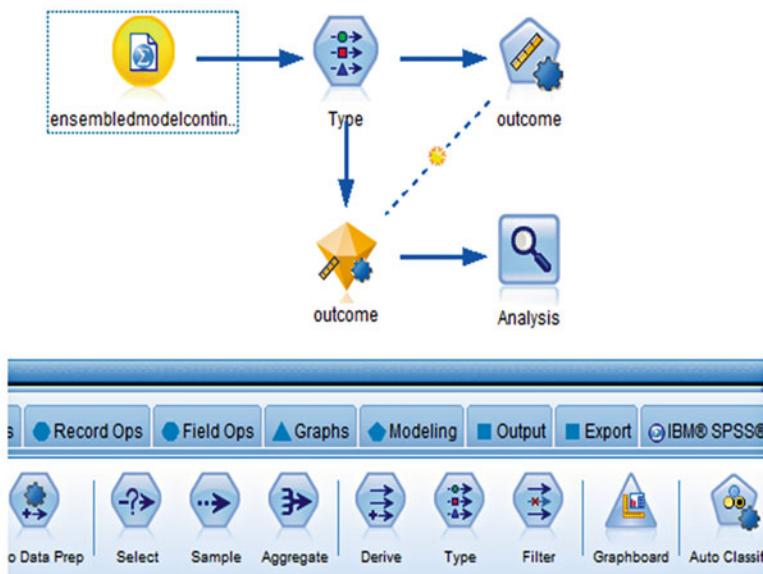
### 17.3 Example

A 250 patients' data file includes 28 variables consistent of patients' gene expression levels and their drug efficacy scores. Only the first 12 patients are shown underneath. The entire data file is in extras.springer.com, and is entitled "spssmodeler1". All of the variables were standardized by scoring them on 11 points linear scales. The following genes were highly expressed: the genes 1–4, 16–19, and 24–27.

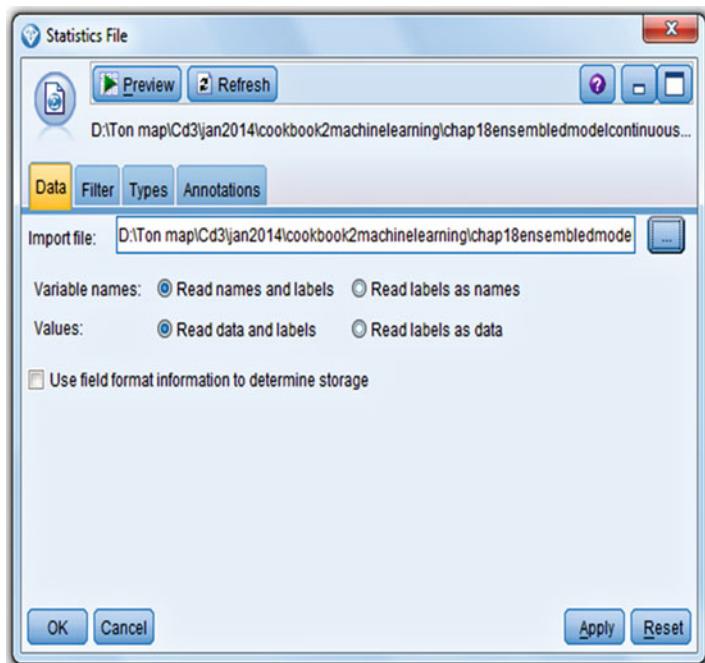
G1	G2	G3	G4	G16	G17	G18	G19	G24	G25	G26	G27	O
8,00	8,00	9,00	5,00	7,00	10,00	5,00	6,00	9,00	9,00	6,00	6,00	7,00
9,00	9,00	10,00	9,00	8,00	8,00	7,00	8,00	8,00	9,00	8,00	8,00	7,00
9,00	8,00	8,00	8,00	8,00	9,00	7,00	8,00	9,00	8,00	9,00	9,00	8,00
8,00	9,00	8,00	9,00	6,00	7,00	6,00	4,00	6,00	6,00	5,00	5,00	7,00
10,00	10,00	8,00	10,00	9,00	10,00	10,00	8,00	8,00	9,00	9,00	9,00	8,00
7,00	8,00	8,00	8,00	8,00	7,00	6,00	5,00	7,00	8,00	8,00	7,00	6,00
5,00	5,00	5,00	5,00	5,00	6,00	4,00	5,00	5,00	6,00	6,00	5,00	5,00
9,00	9,00	9,00	9,00	8,00	8,00	8,00	9,00	8,00	3,00	8,00	8,00	8,00
9,00	8,00	9,00	8,00	9,00	8,00	7,00	7,00	7,00	7,00	5,00	8,00	7,00
10,00	10,00	10,00	10,00	10,00	10,00	10,00	10,00	10,00	8,00	8,00	10,00	10,00
2,00	2,00	8,00	5,00	7,00	8,00	8,00	8,00	9,00	3,00	9,00	8,00	7,00
7,00	8,00	8,00	7,00	8,00	6,00	6,00	7,00	8,00	8,00	8,00	7,00	7,00

G = gene (gene expression levels), O = outcome (score)

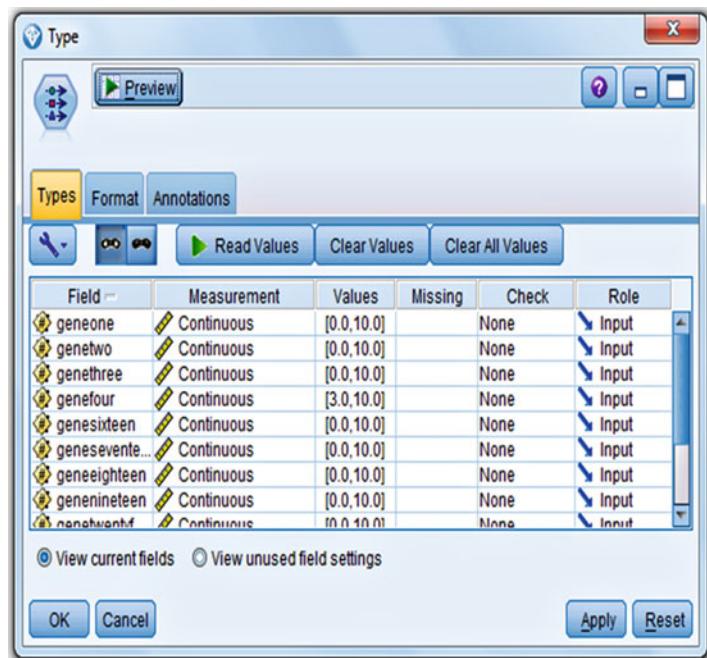
We will start by opening SPSS Modeler version (14.2 and up).



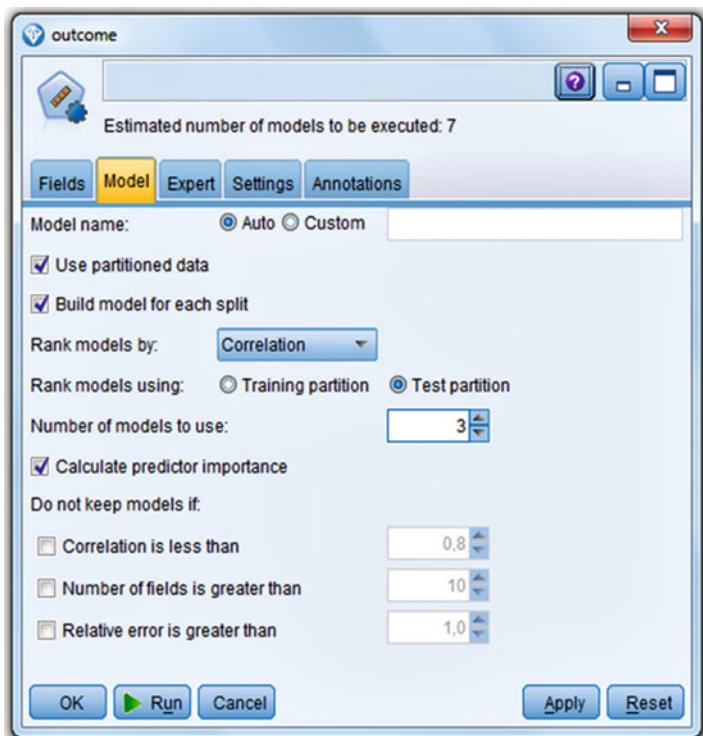
The canvas is, initially, blank, and above a screen view is of the final “completed ensemble” model, otherwise called stream of nodes, which we are going to build. First, in the palettes at the bottom of the screen full of nodes, look and find the **Statistics File node**, and drag it to the canvas. Double-click on it. . .Import file: browse and enter the file “spssmodeler1” . . .click OK. The graph below shows that the data file is open for analysis.



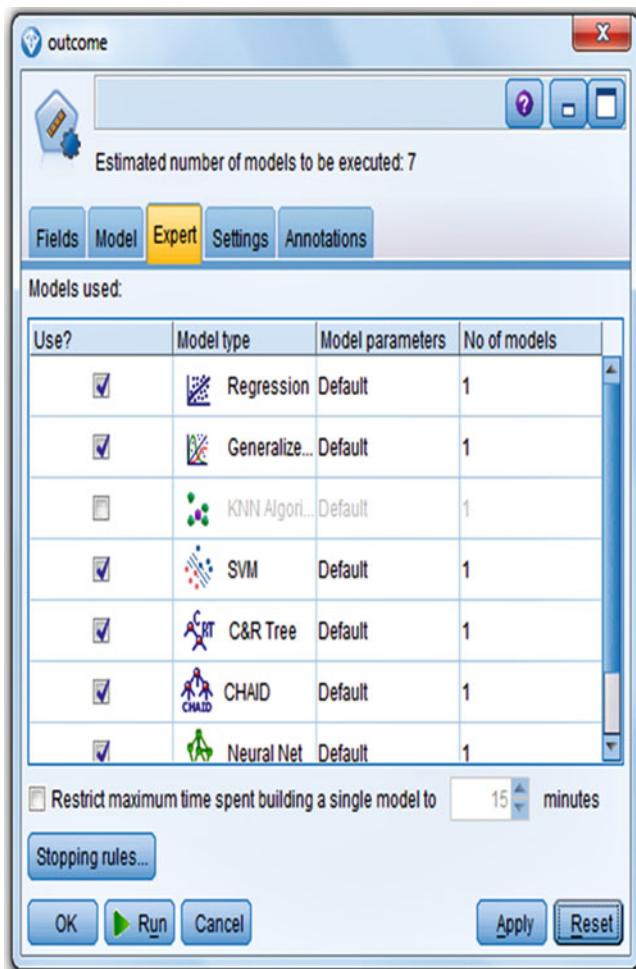
In the palette at the bottom of screen find Type node and drag to the canvas. . .right-click on the Statistics File node. . .a Connect symbol comes up. . .click on the Type node. . .an arrow is displayed. . .double-click on the Type Node. . .after a second or two the underneath graph with information from the Type node is observed. Type nodes are used to access the properties of the variables (often called fields here) like type, role, unit etc. in the data file. As shown below, the variables are appropriately set: 14 predictor variables, 1 outcome (= target) variable, all of them continuous.



Now, click the Auto Numeric node and drag to canvas and connect with the Type node using the above connect-procedure. Click the Auto Numeric node, and the underneath graph comes up....now click Model....select Correlation as metric to rank quality of the various analysis methods used.... the additional manoeuvres are as indicated below....in Numbers of models to use: type the number 3.



Then click the Expert tab. It is shown below. Out of 7 statistical models the three best fit ones are used by SPSS modeler for the ensembled model.



The 7 statistical models include:

- 1/ Linear regression (Regression)
- 2/ Generalized linear model (Generalize...)
- 3/ K nearest neighbor clustering (KNN Algo)
- 4/ Support vector machine (SVM)
- 5/ Classification and regression tree (C&R)
- 6/ Chi square automatic interaction detection (CHAID Tree)
- 7/ Neural network (Neural Net)

More background information of the above methods are available at

- 1/ SPSS for Starters Part One, Chap.5, Linear regression, pp. 15–18, Springer Heidelberg Germany 2010
- 2/ Machine Learning in Medicine a Complete Overview, Chaps. 20 and 21, Generalized linear Models, Springer Heidelberg Germany, 2015, from the same authors.
- 3/ Machine Learning in Medicine a Complete Overview, Chap.1, Springer Heidelberg Germany, 2015, from the same authors.
- 4/ Machine Learning in Medicine Part Two, Chap.15, Support vector machines, pp. 155–161, Springer Heidelberg Germany, 2013.
- 5/ Machine Learning in Medicine a Complete Overview, Chap. 53, Springer Heidelberg Germany, 2015, from the same authors.
- 6/ Machine Learning in Medicine Part Three, Chap.14, Decision trees, pp. 137–150, Springer Heidelberg Germany 2013.
- 7/ Machine Learning in Medicine Part One, Chap.12, Artificial intelligence, multilayer perceptron modeling, pp. 145–154, Springer Heidelberg Germany 2013.

In the above graph click the Settings tab....click the Run button....now a gold nugget is placed on the canvas....click the gold nugget....the model created is shown below.

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
<input checked="" type="checkbox"/>		CHAID 1 <1		0,854	8	0,271
<input checked="" type="checkbox"/>		SVM 1 <1		0,836	12	0,304
<input checked="" type="checkbox"/>		Regressi...<1		0,821	12	0,326

The correlation coefficients of the three best models are over 0.8, and, thus, pretty good. We will now perform the ensembled procedure.

Find in the palettes below the screen the Analysis node and drag it to the canvas. With the above connect procedure connect it with the gold nugget....click the Analysis node.

Comparing \$XR-outcome with outcome

<b>Minimum Error</b>	-2,878
<b>Maximum Error</b>	3,863
<b>Mean Error</b>	-0,014
<b>Mean Absolute Error</b>	0,77
<b>Standard Deviation</b>	1,016
<b>Linear Correlation</b>	0,859
<b>Occurrences</b>	250

The above table is shown and gives the statistics of the ensembled model created. The ensembled outcome is the average score of the scores from the three best fit statistical models. Adjustment for multiple testing and for variance stabilization with Fisher transformation is automatically carried out. The ensembled outcome (named the \$XR-outcome) is compared with the outcomes of the three best fit statistical models, namely, CHAID (chi square automatic interaction detector), SVM (support vector machine), and Regression (linear regression). The ensembled correlation coefficient is larger (0.859) than the correlation coefficients from the three best fit models (0.854, 0.836, 0.821), and so ensembled procedures make sense, because they can provide increased precision in the analysis. The ensembled model can now be stored as an SPSS Modeler Stream file for future use in the appropriate folder of your computer.

## 17.4 Conclusion

In the example given in this chapter, the ensembled correlation coefficient is larger (0.859) than the correlation coefficients from the three best fit separate models (0.854, 0.836, 0.821), and, so, ensembled procedures do make sense, because they can provide increased precision in the analysis.

Ensembled learning is of course different from meta-analysis. With ensembled learning we have multiple analytic models, and a single study. With meta-analysis, we have a single analytic model, and multiple studies. Why do we perform meta-analyses. This is pretty obvious: with more data, we are likely to obtain more certainty. In addition, more differences between subgroups may be observed. Why do we perform ensembled learning. With multiple analytic models, we are likely to obtain again more sensitivity of testing, and, therefore, more certainty of testing. Now, is there a limit to the number of analytic models to be included in ensembled learning procedure. This has not been fully clarified. But, Bonab et al. recently provided data to support that with increasing numbers of models the return on investment rapidly diminishes, the number of models should in any case not exceed the numbers of predictor variables in your model (A theoretical framework on the

ideal number of classifiers for online ensembles in data streams, CIKM USA 2016, p 2053).

In this chapter SPSS modeler was used. It is a software program entirely distinct from SPSS statistical software, though it uses most if not all of the calculus methods of it. It is a standard software package particularly used by market analysts, but, as shown, can, perfectly, well be applied for exploratory purposes in medical research. Alternatively, R statistical software, Knime (Konstanz information miner machine learning software), the packages for Support Vector Machines (LIBSVM), and ensembled support vector machines (ESVM), and many more software programs can be used for ensembled analyses.

## Reference

In this chapter SPSS modeler was used. It is a software program entirely distinct from SPSS statistical software, though it uses most if not all of the calculus methods of it. It is a standard software package particularly used by market analysts, but, as shown, can, perfectly, well be applied for exploratory purposes in medical research. Alternatively, R statistical software, Knime (Konstanz information miner machine learning software), the packages for Support Vector Machines (LIBSVM), and ensembled support vector machines (ESVM), and many more software programs can be used for ensembled analyses.

# **Chapter 18**

## **Ensembled Accuracies**

### **Less Noise but More Risk of Overfit**

**Abstract** Ensemble learning refers to the simultaneous use of multiple learning algorithms for the purpose of a better predictive power. Using SPSS Modeler a 200 patient data file with 11 variables, mainly patients' laboratory values and their subsequent outcome (death or alive), were analyzed with decision trees, logistic regression, Bayesian networks, discriminant analysis, nearest neighbors clustering, support vector machines, chisquare automatic interaction detection, and neural networks. The overall accuracies of the four best fit models were used for computing an average accuracy and its errors. This average accuracy was a lot better than those of the separate algorithms.

### **18.1 Introduction**

The term ensemble learning has been used by many authors from the machine learning community since the early 1990s. It refers to the simultaneous use of multiple learning algorithms for the purpose of obtaining a better predictive potential. Why should a better predictive potential from multiple algorithms be obvious? One may argue that the aggregate of a set of models is less noisy than the result of a single model. However, what about overfit of the aggregate causing statistical tests to be insignificant and inadequate. Mathematically, however, it can be shown that if individual models do not overfit, there will be no room for overfit of the aggregate. In addition it can be argued that often different models as used have their own way of classifying data. In line with Occam, the English Franciscan monk from the thirteenth century, traditional statistics says: the simplest model provides the best power. However, in line with Epicurus, the philosopher from Athens 300 BC, the machine learning community started to doubt the traditional concept, and started to think differently: if multiple analytic models explain data, keep them all.

Regarding the argument that different models as used have their own way of classifying data, an example will be given. In the data example of the next paragraph 11 variables consistent of patients' laboratory values and their subsequent outcome (death or alive). SPSS Modeler (see also Chap. 48, Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors) was used to select the best fit models and perform an ensembled analysis. The four best fit models were

- (1) Bayesian networks (Chap. 70, Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors)
- (2) KNN (k nearest neighbors algorithm) (Chap. 4, Machine learning inmedicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors)
- (3) Logistic regression (Chap. 36, SPSS for starters and 2nd levelers, Springer Heidelberg Germany, 2016, from the same authors)
- (4) Neural networks (Chap. 50, Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2015, from the same authors)

Unlike regression models, Bayesian networks have predictors that are independent of one another. They are not for assessing strictly associations but rather causal relations, and make use of path statistics and directed acyclic graphs. KNN is a cluster method requiring a lot of computer calculus, because the distance of one value to all of the other values in a data file is measured. Logistic regression is like linear regression but with a binary outcome variable. Unlike regression analyses, which use algebraic functions for data fitting, neural networks uses a stepwise method called the steepest descent method, in order to assess whether typically nonlinear relationships can adequately fit the data.

The four methodologies selected by the computer, have, obviously, ways of classifying data that are far from similar. This is, furthermore, supported by the example given underneath, that will show that largely different numbers of analysis steps, here called fields, and different outcome values and error measures are produced.

## 18.2 Ensemble Learning with SPSS Modeler

SPSS modeler is a work bench for automatic data mining (Chap. 61, Machine learning in medicine, Springer Heidelberg Germany, 2015, from the same authors) and modeling (see also the Chaps. 64 and 65, Machine learning in medicine, Springer Heidelberg Germany, 2015, from the same authors). So far it is virtually unused in medicine, and mainly applied by econo-/sociometrists. Automatic modeling of binary outcomes computes the ensembled result of a number of best fit models for a particular data set, and provides better sensitivity than the separate models do. This chapter is to demonstrate its performance with clinical event prediction. Multiple laboratory values can predict events like health, death, morbidities etc. Can ensembled modeling with four best fit statistical models provide better precision than the separate analysis with single statistical models does.

## 18.3 Example

A 200 patients' data file includes 11 variables consistent of patients' laboratory values and their subsequent outcome (death or alive). Only the first 12 patients are shown underneath. The entire data file is in extras.[springer.com](http://springer.com), and is entitled "spssmodeler2".

Death	ggt	asat	alat	bili	ureum	creat	c-clear	esr	crp	leucos
,00	20,00	23,00	34,00	2,00	3,40	89,00	-111,00	2,00	2,00	5,00
,00	14,00	21,00	33,00	3,00	2,00	67,00	-112,00	7,00	3,00	6,00
,00	30,00	35,00	32,00	4,00	5,60	58,00	-116,00	8,00	4,00	4,00
,00	35,00	34,00	40,00	4,00	6,00	76,00	-110,00	6,00	5,00	7,00
,00	23,00	33,00	22,00	4,00	6,10	95,00	-120,00	9,00	6,00	6,00
,00	26,00	31,00	24,00	3,00	5,40	78,00	-132,00	8,00	4,00	8,00
,00	15,00	29,00	26,00	2,00	5,30	47,00	-120,00	12,00	5,00	5,00
,00	13,00	26,00	24,00	1,00	6,30	65,00	-132,00	13,00	6,00	6,00
,00	26,00	27,00	27,00	4,00	6,00	97,00	-112,00	14,00	6,00	7,00
,00	34,00	25,00	13,00	3,00	4,00	67,00	-125,00	15,00	7,00	6,00
,00	32,00	26,00	24,00	3,00	3,60	58,00	-110,00	13,00	8,00	6,00
,00	21,00	13,00	15,00	3,00	3,60	69,00	-102,00	12,00	2,00	4,00

death = death yes no (0 = no)

ggt = gamma glutamyl transferase (u/l)

asat = aspartate aminotransferase (u/l)

alat = alanine aminotransferase (u/l)

bili = bilirubine (micromol/l)

ureum = ureum (mmol/l)

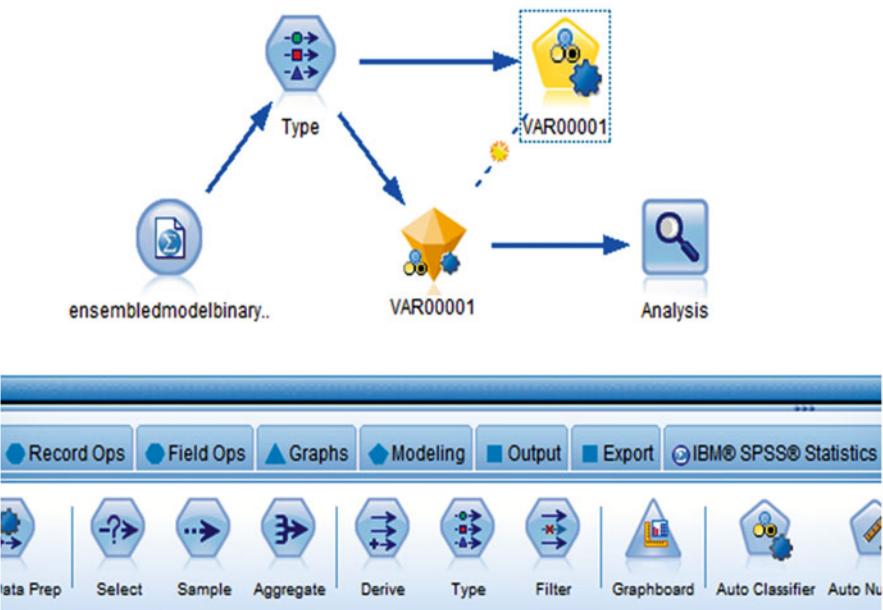
creat = creatinine (mmicromol/l)

c-clear = creatinine clearance (ml/min)

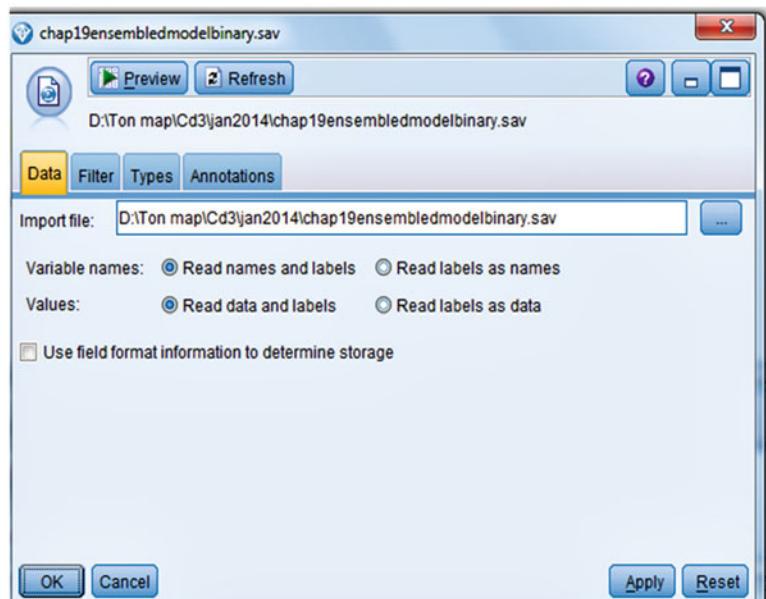
esr = erythrocyte sedimentation rate (mm)

crp = c-reactive protein (mg/l)

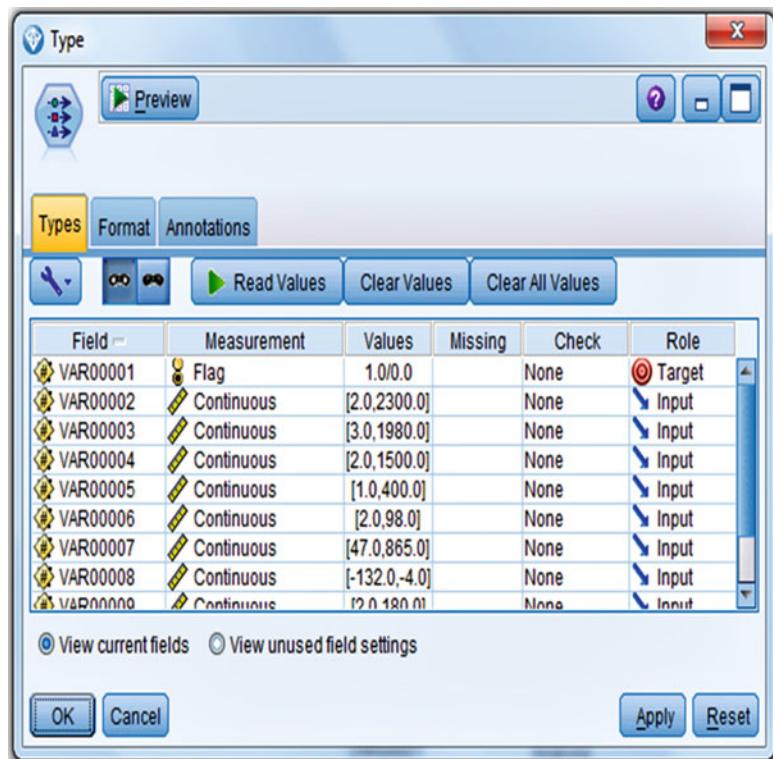
leucos = leucocyte count (.10<sup>9</sup>/l)



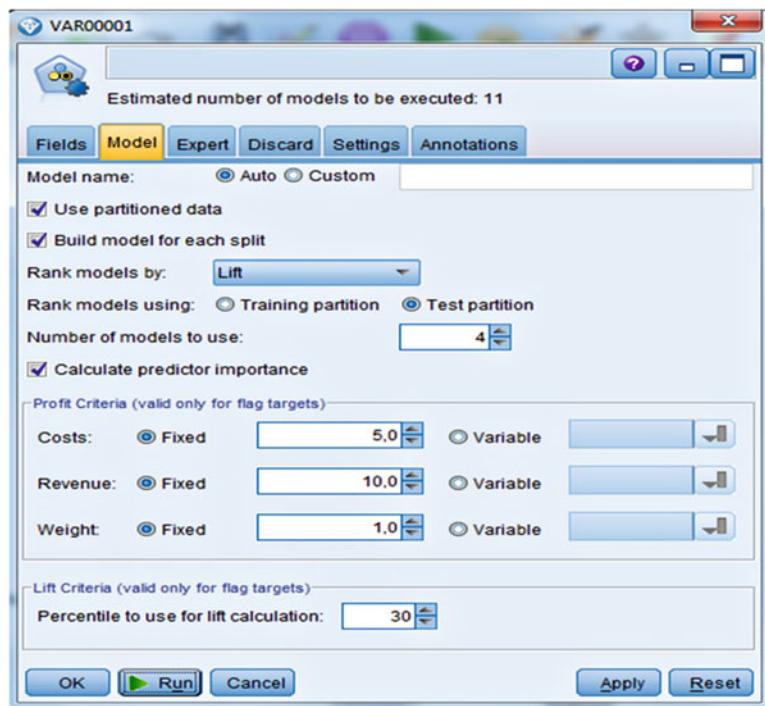
The canvas is, initially, blank, and above is given a screen view of the completed ensembled model, otherwise called stream of nodes, which we are going to build. First, in the palettes at the bottom of the screen full of nodes, look and find the **Statistics File node**, and drag it to the canvas, pressing the mouse left side. Double-click on this node....Import file: browse and enter the file “chap19ensembledmodelbinary” .... click OK. The graph below shows, that the data file is open for analysis.



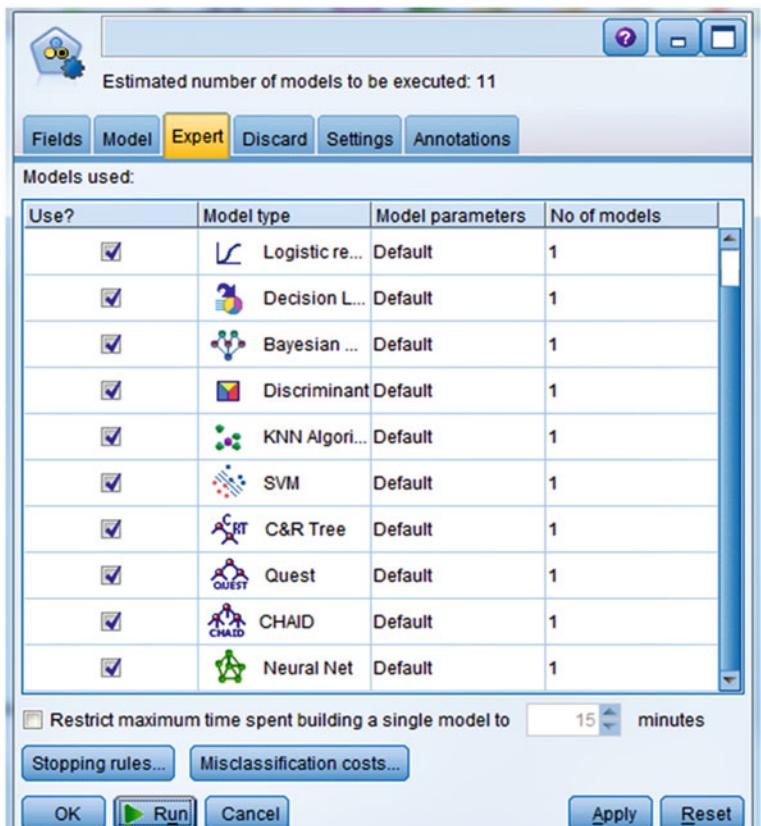
In the palette at the bottom of screen find Type node and drag to the canvas. . . . right-click on the Statistics File node. . . . a Connect symbol comes up. . . . click on the Type node. . . . an arrow is displayed. . . . double-click on the Type Node. . . . after a second or two the underneath graph with information from the Type node is observed. Type nodes are used to access the properties of the variables (often called fields here) like type, role, unit etc. in the data file. As shown below, 10 predictor variables (all of them continuous) are appropriately set. However, VAR 00001 (death) is the outcome (= target) variable, and is binary. Click in the row of variable VAR00001 on the measurement column and replace "Continuous" with "Flag". Click Apply and OK. The underneath figure is removed and the canvas is displayed again.



Now, click the Auto Classifier node and drag to the canvas, and connect with the Type node using the above connect-procedure. Click the Auto Classifier node, and the underneath graph comes up....now click Model....select Lift as Rank model of the various analysis models used.... the additional manoeuvres are as indicated below....in Numbers of models to use: type the number 4.



Then click the Expert tab. It is shown below. Out of 11 statistical models the four best fit ones are selected by SPSS modeler for constructing an ensembled model.



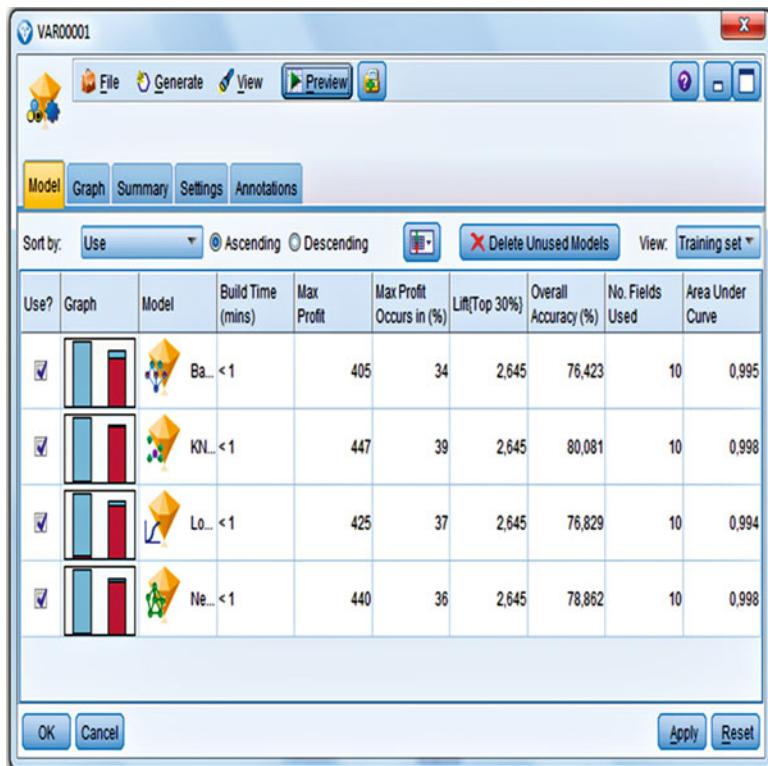
The 11 statistical analysis methods for a flag target (= binary outcome) include:

- 1/ C5.0 decision tree (C5.0)
- 2/ Logistic regression (Logist re...)
- 3/ Decision list (Decision L....)
- 4/ Bayesian network (Bayesian. ....)
- 5/ Discriminant analysis (Discriminant)
- 6/ K nearest neighbors algorithm (KNN Algori...)
- 7/ Support vector machine (SVM)
- 8/ Classification and regression tree (C&R Tree)
- 9/ Quest decision tree (Quest. ....)
- 10/ Chi square automatic interaction detection (CHAID)
- 11/ Neural network (Neural Net)

More background information of the above methods are available at.

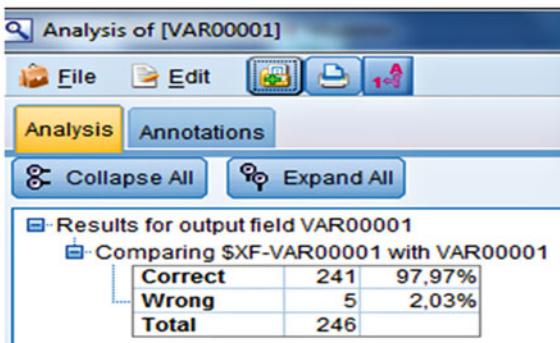
- 1/ Chap. 24 of current work (Automatic data mining in SPSS Modeler).
- 2/ SPSS for Starters Part One, Chap. 11, Logistic regression, pp. 39–42, Springer Heidelberg Germany 2010.
- 3/ Decision list models identify high and low performing segments in a data file,
- 4/ Machine Learning in Medicine Part Two, Chap. 16, Bayesian networks, pp. 163–170, Springer Heidelberg Germany, 2013.
- 5/ Machine Learning in Medicine Part One, Chap. 17, Discriminant analysis for supervised data, pp. 215–224, Springer Heidelberg Germany 2013.
- 6/ Machine Learning in Medicine a Complete Overview, Chap. 4, Nearest neighbors for classifying new medicines, Springer Heidelberg Germany, 2015, from the same authors.
- 7/ Machine Learning in Medicine Part Two, Chap. 15, Support vector machines, pp. 155–161, Springer Heidelberg Germany, 2013.
- 8/ Machine Learning in Medicine, Chap. 53, Decision trees for decision analysis, Springer Heidelberg Germany, 2015, from the same authors.
- 9/ QUEST (Quick Unbiased Efficient Statistical Trees) are improved decision trees for binary outcomes.
- 10/ Machine Learning in Medicine Part Three, Chap. 14, Decision trees, pp. 137–150, Springer Heidelberg Germany 2013.
- 11/ Machine Learning in Medicine Part One, Chap. 12, Artificial intelligence, multilayer perceptron modeling, pp. 145–154, Springer Heidelberg Germany 2013.

In the above graph click the Settings tab....click the Run button....now a gold nugget is placed on the canvas....click the gold nugget....the model created is shown below.



The overall accuracies (%) of the four best fit models are over 0.8, and are, thus, pretty good. We will now perform the ensembled procedure.

Find in the palettes at the bottom of the screen the Analysis node and drag it to the canvas. With above connect procedure connect it with the gold nugget. . .click the Analysis node.



The above table is shown and gives the statistics of the ensembled model created. The ensembled outcome is the average accuracy of the accuracies from the four best fit statistical models. In order to prevent overstated certainty due to overfitting, bootstrap aggregating (“bagging”) is used. The ensembled outcome (named the \$XR-outcome) is compared with the outcomes of the four best fit statistical models, namely, Bayesian network, k Nearest Neighbor clustering, Logistic regression, and Neural network. The ensembled accuracy (97.97%) is much larger than the accuracies of the four best fit models (76.423, 80.081, 76.829, and 78.862%), and, so, ensembled procedures make sense, because they provide increased precision in the analysis. The computed ensembled model can now be stored in your computer in the form of an SPSS Modeler Stream file for future use.

## 18.4 Conclusion

In the example given in this chapter, the ensembled accuracy is larger (97.97%) than the accuracies from the four best fit models (76.423, 80.081, 76.829, and 78.862%), and so ensembled procedures make sense, because they can provide increased precision in the analysis.

Ensembled learning is of course different from meta-analysis. With ensembled learning we have multiple analytic models, and a single study. With meta-analysis, we have a single analytic model, and multiple studies. Why do we perform meta-analyses. This is pretty obvious: with more data, we are likely to obtain more certainty. In addition, more differences between subgroups may be observed. Why do we perform ensembled learning. With multiple analytic models, we are likely to obtain again more sensitivity of testing, and, therefore, more certainty of testing. Now, is there a limit to the number of analytic models to be included in ensembled learning procedure. This has not been fully clarified. But, Bonab et al. recently provided data to support that with increasing numbers of models the return on investment rapidly diminishes, the number of models should in any case not exceed the numbers of predictor variables in your model (A theoretical framework on the ideal number of classifiers for online ensembles in data streams, CIKM USA 2016, p 2053).

In this chapter SPSS modeler was used. It is a software program entirely distinct from SPSS statistical software, though it uses most if not all of the calculus methods of it. It is a standard software package particularly used by market analysts, but, as shown, can, perfectly, well be applied for exploratory purposes in medical research. Alternatively, R statistical software, Knime (Konstanz information miner machine learning software), the packages for Support Vector Machines (LIBSVM), and ensembled support vector machines (ESVM), and many more software programs can be used for ensembled analyses.

## Reference

In this chapter SPSS modeler, a work bench for automatic data mining and data modeling from IBM, was used. It is an analytics software application entirely distinct from SPSS statistical software, though it uses most if not all of the calculus methods of it. It is a standard software package particularly used by market analysts, but, as shown, can, perfectly, well be applied for exploratory purposes in medical research. Alternatively, R statistical software, Knime (Konstanz information miner machine learning software), the packages for Support Vector Machines (LIBSVM), and ensembled support vector machines (ESVM), and many more software programs can be used for ensembled analyses.

# **Chapter 19**

## **Multivariate Meta-analysis**

### **Assessing Separate Effects of Predictors on One Outcome Adjusted for the Other**

**Abstract** Multivariate analysis, simultaneously, assesses the separate effects of the predictors on one outcome adjusted for the other. E.g., it can answer clinically important questions like: does drug-compliance not only predict drug efficacy, but also, independently of the first effect, predict quality of life (qol). In this chapter (1) the effect of counseling and non-compliance on drug efficacy and quality of life was assessed in a meta-analysis of 25 studies, (2) the effect of type of research group on hospital admissions due to adverse drug effects, study magnitudes, and patients age, in a meta-analysis of 20 studies, and (3) the effect of counseling and non-compliance on drug efficacy and quality of life in a meta-analysis of 20 studies. In all of the examples a beneficial effect of the predictors on the multivariate outcomes was observed.

#### **19.1 Introduction**

Multivariate analysis is a method that, simultaneously, assesses more than a single outcome variable. It is different from repeated measures analysis of variance and mixed models, that assess both the difference between the outcomes and the overall effects of the predictors on the outcomes. Multivariate analysis, simultaneously, assesses the separate effects of the predictors on one outcome adjusted for the other. E.g., it can answer clinically important questions like: does drug-compliance not only predict drug efficacy, but also, independently of the first effect, predict quality of life (qol). Path statistics can be used as an approach to multivariate analysis of variance (MANOVA) (see SPSS for starters and second levelers 2nd edition, 2016, Chap. 17, Springer Heidelberg Germany, from the same authors). However, MANOVA is the real thing, because it produces an overall level of significance of a predictive model with multiple outcome and predictor variables. Meta-analyses of multivariate diagnostic studies have been published by members of our group (Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; 58: 982–90) with log odds ratios and their standard errors and log odds transformed values of sensitivities and “1-specificities” and their standard errors. In this chapter three examples are given of multivariate meta-analyses of therapeutic studies. In example 1 point estimates of the studies without measure of spread were meta-analyzed, in example 2 study sample sizes were included in the analysis as predictor variable, in example 3 study sample sizes were used for data weighting purposes instead of standard errors.

## 19.2 Example 1

25 Studies of the effects of counseling and non-compliance on drug efficacy and quality of life were meta-analyzed. Drug efficacy was measured as mean stools per month, quality of life (qol) as mean qol scores per study, counselings as mean counsellings per study and compliance as mean non-compliance scores with drug treatment per study.

Outcome 1	outcome 2	predictor 1	predictor 2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

outcome 1 = mean stools per month

outcome 2 = mean qol scores per study

predictor 1 = mean counsellings per study

predictor 2 = mean non-compliance score per study

The primary question of the studies was: do two outcome variables enable to make better use of predicting variables than a single one.

The mean results of the first 10 studies of the 25 study data file is given underneath.

stools	qol	counsel	compliance
41,00	112,00	11,00	36,00
38,00	99,00	11,00	30,00
39,00	86,00	12,00	27,00
37,00	107,00	10,00	38,00
47,00	108,00	18,00	40,00
30,00	95,00	13,00	31,00
36,00	88,00	12,00	25,00
12,00	67,00	4,00	24,00
26,00	112,00	10,00	27,00
20,00	87,00	8,00	20,00

stools = mean stools per month

qol = mean quality of life scores per study

counseling = mean counsellings per study

compliance = mean non-compliance with drug treatment per study

The entire data file is entitled “multivariate”, and is in [extras.springer.com](http://extras.springer.com). Start by opening the data file in SPSS. The module General Linear Model consists of 4 statistical models:

Univariate,

Multivariate,

Repeated Measures,

Variance Components.

We will use here the statistical model Multivariate.

We will first assess whether counseling frequency is a significant predictor of (1) both frequency improvement of stools and (2) improved quality of life.

### Command

Analyze...General Linear Model...Multivariate...In dialog box Multivariate: transfer "therapeutic efficacy" and "qol" to Dependent Variables and "counseling and non-compliance" to Fixed factors ...OK.

Multivariate Tests<sup>a</sup>

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	,997	303,329 <sup>b</sup>	1,000	1,000	,037
	Wilks' Lambda	,003	303,329 <sup>b</sup>	1,000	1,000	,037
	Hotelling's Trace	303,329	303,329 <sup>b</sup>	1,000	1,000	,037
	Roy's Largest Root	303,329	303,329 <sup>b</sup>	1,000	1,000	,037
counseling	Pillai's Trace	,924	1,359 <sup>b</sup>	9,000	1,000	,587
	Wilks' Lambda	,076	1,359 <sup>b</sup>	9,000	1,000	,587
	Hotelling's Trace	12,231	1,359 <sup>b</sup>	9,000	1,000	,587
	Roy's Largest Root	12,231	1,359 <sup>b</sup>	9,000	1,000	,587
compliance	Pillai's Trace	,764	,361 <sup>b</sup>	9,000	1,000	,870
	Wilks' Lambda	,236	,361 <sup>b</sup>	9,000	1,000	,870
	Hotelling's Trace	3,245	,361 <sup>b</sup>	9,000	1,000	,870
	Roy's Largest Root	3,245	,361 <sup>b</sup>	9,000	1,000	,870
counseling * compliance	Pillai's Trace	,000	<sup>b</sup>	,000	,000	.
	Wilks' Lambda	1,000	<sup>b</sup>	,000	1,000	.
	Hotelling's Trace	,000	<sup>b</sup>	,000	2,000	.
	Roy's Largest Root	,000	,000 <sup>b</sup>	1,000	,000	.

a. Design: Intercept + counseling + compliance + counseling \* compliance

b. Exact statistic

The above table shows that MANOVA can be considered as another regression model with intercepts and regression coefficients. Just like analysis of variance (ANOVA) it is based on normal distributions and homogeneity of the variables. Neither counseling, nor compliance, nor their interaction were significant predictors of the bivariate outcome (1) drug efficacy and (2) quality of life.

Next, we will perform a MANOVA with a single predictor variable and the same bivariate outcome (1) drug efficacy and (2) quality of life. First the effect of counseling will be tested.

### Command

Analyze...General Linear Model...Multivariate...In dialog box Multivariate: transfer "therapeutic efficacy" and "qol" to Dependent Variables and "counseling" to Fixed factors ....OK.

**Multivariate Tests<sup>a</sup>**

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	,996	1007,512 <sup>b</sup>	2,000	9,000	,000
	Wilks' Lambda	,004	1007,512 <sup>b</sup>	2,000	9,000	,000
	Hotelling's Trace	223,892	1007,512 <sup>b</sup>	2,000	9,000	,000
	Roy's Largest Root	223,892	1007,512 <sup>b</sup>	2,000	9,000	,000
counseling	Pillai's Trace	1,571	2,619	28,000	20,000	,015
	Wilks' Lambda	,029	3,165 <sup>b</sup>	28,000	18,000	,007
	Hotelling's Trace	13,033	3,724	28,000	16,000	,004
	Roy's Largest Root	11,145	7,961 <sup>c</sup>	14,000	10,000	,001

a. Design: Intercept + counseling

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

The above table shows that counseling is a very significant predictor of the bivariate outcome.

**Tests of Between-Subjects Effects**

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	therapeutic efficacy	2418,973 <sup>a</sup>	14	172,784	5,653	,005
	qol	5538,527 <sup>b</sup>	14	395,609	3,681	,022
Intercept	therapeutic efficacy	19780,042	1	19780,042	647,112	,000
	qol	171915,690	1	171915,690	1599,464	,000
counseling	therapeutic efficacy	2418,973	14	172,784	5,653	,005
	qol	5538,527	14	395,609	3,681	,022
Error	therapeutic efficacy	305,667	10	30,567		
	qol	1074,833	10	107,483		
Total	therapeutic efficacy	26564,000	25			
	qol	212911,000	25			
Corrected Total	therapeutic efficacy	2724,640	24			
	qol	6613,360	24			

a. R Squared = ,888 (Adjusted R Squared = ,731)

b. R Squared = ,837 (Adjusted R Squared = ,610)

The above table is also in the output, and gives separate levels of significance of the effect of counseling on the two outcomes, which are pretty good, 0.005 and 0.022.

Next, we will perform a MANOVA with a single predictor variable and the same bivariate outcome (1) drug efficacy and (2) quality of life. The effect of non-compliance will be tested.

### Command

Analyze...General Linear Model...Multivariate...In dialog box Multivariate: transfer "therapeutic efficacy" and "qol" to Dependent Variables and "compliance" to Fixed factors ...OK.

Multivariate Tests<sup>a</sup>

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	,991	483,239 <sup>b</sup>	2,000	9,000	,000
	Wilks' Lambda	,009	483,239 <sup>b</sup>	2,000	9,000	,000
	Hotelling's Trace	107,387	483,239 <sup>b</sup>	2,000	9,000	,000
	Roy's Largest Root	107,387	483,239 <sup>b</sup>	2,000	9,000	,000
compliance	Pillai's Trace	1,337	1,441	28,000	20,000	,201
	Wilks' Lambda	,103	1,362 <sup>b</sup>	28,000	18,000	,250
	Hotelling's Trace	4,443	1,269	28,000	16,000	,314
	Roy's Largest Root	3,031	2,165 <sup>c</sup>	14,000	10,000	,112

a. Design: Intercept + compliance

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

The above table shows that according to Roy's test a trend to significance is observed.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	therapeutic efficacy	1771,973 <sup>a</sup>	14	126,570	1,329	,331
	qol	4971,693 <sup>b</sup>	14	355,121	2,163	,112
Intercept	therapeutic efficacy	21545,424	1	21545,424	226,159	,000
	qol	176043,010	1	176043,010	1072,343	,000
compliance	therapeutic efficacy	1771,973	14	126,570	1,329	,331
	qol	4971,693	14	355,121	2,163	,112
Error	therapeutic efficacy	952,667	10	95,267		
	qol	1641,667	10	164,167		
Total	therapeutic efficacy	26564,000	25			
	qol	212911,000	25			
Corrected Total	therapeutic efficacy	2724,640	24			
	qol	6613,360	24			

a. R Squared = ,650 (Adjusted R Squared = ,161)

b. R Squared = ,752 (Adjusted R Squared = ,404)

The above table shows the same: compliance tends to predict the bivariate outcome at p = 0.112. The analysis of this model stops here.

### 19.3 Example 2

Table 1 shows the individual studies included in a meta-analysis of 20 studies on adverse drug admissions (ADEs) of our study group (Atiqi, Cleophas, Van Bommel, Zwinderman, Meta-analysis of recent studies of patients admitted for adverse drug effects, Int J Clin Pharmacol Ther 2009; 47: 549–56). The pooled result of the 20 studies as included in the meta-analysis provided an overall percentage of ADEs of 5.4% (5.0–5.8). However, the meaning of this pooled result was limited due to a significant heterogeneity between the individual studies: both the fixed effects and random effect tests for heterogeneity were highly significant (both  $p < 0.001$ ,  $I^2 > 90\%$ , see the Chaps. 1 and 2 for explanation of the terms). Study characteristics are given underneath.

Study	study size	percentage of all admissions	95% confidence intervals
1.Mannesse et al 2000	106	21.0 %	13.0-29.0 %
2. Malhotra et al 2001	578	14.4 %	11.5-17.3 %
3. Chan et al 2001	240	30.4 %	24.5-36.3 %
4. Olivier et al 2002	671	6.1 %	4.3-7.9 %
5. Mjorndal et al 2002	681	12.0 %	9.5-14.5 %
6. Onder et al. 2002	28411	3.4 %	3.2-3.6 %
7. Koh et al 2003	347	6.6 %	4.0-9.2 %
8. Easton-Carter et al 2003	8601	3.3 %	2.9-3.7 %
9. Dormann et al 2003	915	4.9 %	4.9-14.3 %
10.Peyriere et al 2003	156	9.6 %	4.9-14.3 %
11.Howard et al	4093	6.5 %	5.7-7.3 %
12.Pirmohamed et al	18820	6.5 %	6.2-6.8 %
13.Hardmeier et al 2004	6383	4.1 %	3.6-4.6 %
14.Easton et al 2004	2933	4.3 %	3.6-5.0 %
15.Capuano et al. 2004	480	3.5 %	1.9-5.1 %
16.Caamano et al 2005	19070	4.3 %	3.7-4.6 %
17.Yee et al 2005	2169	12.6 %	11.2-14.0 %
18.Baena et al 2006	2261	33.2 %	31.2-35.2 %
19.Leendertse et al 2006	12793	5.6 %	5.2-6.0 %
20.Van der Hooft et al 2008	355	5.1 %	2.8-7.4 %
Pooled	113203	5.4 %	5.0-5.8 %

A data file suitable for a meta-regression (ADE = admissions due to adverse drug effects) is given underneath.

Study No	%ADEs	Study magnitude	Clinicians' study yes = 1	Elderly study yes = 1
1	21,00	106,00	1,00	1,00
2	14,40	578,00	1,00	1,00
3	30,40	240,00	1,00	1,00
4	6,10	671,00	0,00	0,00
5	12,00	681,00	0,00	0,00
6	3,40	28411,00	1,00	0,00
7	6,60	347,00	0,00	0,00
8	3,30	8601,00	0,00	0,00
9	4,90	915,00	0,00	0,00
10	9,60	156,00	0,00	0,00
11	6,50	4093,00	0,00	0,00
12	6,50	18820,00	0,00	0,00
13	4,10	6383,00	0,00	0,00
14	4,30	2933,00	0,00	0,00
15	3,50	480,00	0,00	0,00
16	4,30	19070,00	1,00	0,00
17	12,60	2169,00	1,00	0,00
18	33,20	2261,00	0,00	1,00
19	5,60	12793,00	0,00	0,00
20	5,10	355,00	0,00	0,00

Studies were also assessed for type-of-research-group. Because medical articles are often co-authored by specialists from different disciplines and/or guest-authors, it was decided to name the type-of-research-group after the type of department of the first two authors.

The studies 1–3 and 18 were performed by clinicians. In them the percentages ADEs were much higher (pooled data 29.2%, 26.4–32.0) than they were in the other studies (pooled data 4.8%, 4.1–5.5). A careful further examination of these data revealed that the difference between the type of research groups was associated with a significant difference in magnitude of the studies: the four clinicians' studies had a mean sample size of 796 (740–852), while the remainder of the studies averaged at 6680 (6516–6844) patients per study, different at  $p < 0.001$ . This effect was ascribed to the presence of publication bias in this meta-analysis (small studies with small results were under-published).

In order to simultaneously assess the effects of study-magnitude, patients' age, and type-of- research-group a multiple linear regression was, subsequently, performed. After adjustment for patients' age and type-of-research-group the study-magnitude was no significant predictor of study effect anymore. In contrast,

the type-of-research-group was the single and highly significant predictor of study result.

SPSS ([www.spss.com](http://www.spss.com)) the underneath multiple linear meta – regression table with study result (percentage ADEs) as dependent variable and study-magnitude, patients' age, and type-of-research-group as independent variables. After adjustment for patients' age and type-of-research-group the study-magnitude is no significant predictor of study effect anymore. In contrast, the type-of-research-group is the single highly-significant predictor of the study result.

Covariate	Unstandardized coefficients		Standardized coefficients		
	B	std error	Beta	t	sig.
Constant	6.92	1.45		4.76	0.000
Study magnitude	-7.7.e-0.05	0.00	-0.071	-0.50	0.62
Patients' age	-1.39	2.89	-0.075	-0.48	0.64
Type research group	18.93	3.36	0.89	5.64	0.000

Dependent variable: study result; std error = standard error; t = t-value;  
sig. = level of significance.

So far, the conclusion of the above meta-analysis was that the type of research group was the single and highly relevant predictor of heterogeneity between 20 studies of on adverse drug admissions. The clinicians included many more adverse drug admissions (ADEs) than did the pharmacists: 29.2% versus 4.8%. However, in a multiple linear regression with percentages of ADEs as outcome and both study magnitude and age class as predictors, both predictors were (borderline) statistically significant at 0.037 and 0.040. This is shown below.

Model	Coefficients <sup>a</sup>					
	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta			
1	(Constant)	10,100	2,247		4,495	,000
	study-magnitude	-,001	,000	-,466	-2,264	,037
	elderly=1	8,515	3,833	,457	2,221	,040

a. Dependent Variable: percentageADEs

In order to assess whether the type of research group might predict multiple factors, including percentage ADEs, study magnitude and age class, multivariate analyses were performed in SPSS statistical software.

Commands, identical to those described in the section entitled Example 1 of this chapter, produced the underneath results.

#### Multivariate Tests<sup>a</sup>

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	,919	60,191 <sup>b</sup>	3,000	16,000	,000
	Wilks' Lambda	,081	60,191 <sup>b</sup>	3,000	16,000	,000
	Hotelling's Trace	11,286	60,191 <sup>b</sup>	3,000	16,000	,000
	Roy's Largest Root	11,286	60,191 <sup>b</sup>	3,000	16,000	,000
investigatortype	Pillai's Trace	,808	22,491 <sup>b</sup>	3,000	16,000	,000
	Wilks' Lambda	,192	22,491 <sup>b</sup>	3,000	16,000	,000
	Hotelling's Trace	4,217	22,491 <sup>b</sup>	3,000	16,000	,000
	Roy's Largest Root	4,217	22,491 <sup>b</sup>	3,000	16,000	,000

a. Design: Intercept + investigatortype

b. Exact statistic

The above overall analysis with one predictor (type of research group) and three outcomes (1 ADEs, 2 study magnitude, 3 age class) showed that all of the four models used by SPSS were very significant at  $p < 0.0001$ . The type of research group did not only predict the ADEs, but also the combination of the three outcomes 1 ADEs 2 study magnitude 3 age class and very significantly so. In order to identify which of the three outcomes was the most important one SPSS also provided in the output sheets a between-subject MANOVA.

#### Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	percentageADEs	1107,072 <sup>a</sup>	1	1107,072	57,031	,000
	study-magnitude	110774538,... <sup>b</sup>	1	110774538,...	1,773	,200
	elderly=1	1,013 <sup>c</sup>	1	1,013	5,718	,028
Intercept	percentageADEs	3055,392	1	3055,392	157,400	,000
	study-magnitude	178855824,...	1	178855824,...	2,862	,108
	elderly=1	2,813	1	2,813	15,882	,001
investigatortype	percentageADEs	1107,072	1	1107,072	57,031	,000
	study-magnitude	110774538,...	1	110774538,...	1,773	,200
	elderly=1	1,013	1	1,013	5,718	,028

The above table shows, that the type of research group (investigatortype) significantly predicted ADEs and age class (here named elderly = 1) at  $p = 0.0001$  and 0.28, and it gave a trend for study magnitude at  $p = 0.200$ .

Additionally, a bivariate MANOVA was provided with ADEs and age class as dependent variables only.

**Tests of Between-Subjects Effects**

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	percentageADEs	1107,072 <sup>a</sup>	1	1107,072	57,031	,000
	elderly=1	1,013 <sup>b</sup>	1	1,013	5,718	,028
Intercept	percentageADEs	3055,392	1	3055,392	157,400	,000
	elderly=1	2,813	1	2,813	15,882	,001
investigatortype	percentageADEs	1107,072	1	1107,072	57,031	,000
	elderly=1	1,013	1	1,013	5,718	,028

Again significant predictor effects were observed of the predictor investigatortype, were observed, and they were so at the same level of significance as the three outcomes MANOVA.

And so, the ultimate conclusions of the meta-analyses had to be slightly adjusted after the meta-analytic MANOVAS. Both ADEs and age class were significant outcomes of investigatortype, ans study magnitude provided a trend to significance to the same bivariate outcomes at  $p = 0.20$ .

## 19.4 Example 3

In a meta-analysis of 20 studies the outcome measures were the proportion of patients with a high efficacy drug response and the proportion of patients with a high quality of life score. The predictors variables were the numbers of counseling sessions per patient per protocol, and the numbers of noncompliance events per study. Also the study sizes were included. The file of the meta-data is given below.

Variable		1	2	3	4	5
14	75	5	18	90		
39	99	13	14	89		
42	100	15	30	120		
41	98	11	36	77		
38	99	11	30	45		
39	86	12	27	78		
37	77	10	38	67		
47	97	18	40	100		
30	95	13	31	95		
36	88	12	25	78		
12	67	4	24	79		
26	76	10	27	98		
20	87	8	20	99		
43	100	16	35	96		
31	93	15	29	67		
40	92	14	32	83		
31	78	7	30	99		
36	100	12	40	95		
21	69	6	31	92		
44	66	19	41	90		

Variable 1 = proportion of patients with high efficacy drug response

Variable 2 = proportion of patients with high quality of life score

Variable 3 = number of counseling sessions per patient per protocol

Variable 4 = number of noncompliance events per study

First, we will weigh the variables by the sample size of the studies according to:

sample size	study 1	study 2	study 3	
	.....			
				+
summary of sample sizes				
weighting factor = (summary of sample sizes) / 20				
weighting factor	1	2	3	4
	weighted			
1,036	14,50	77,70	5,18	18,65
1,025	39,97	101,48	13,33	14,35
1,382	58,04	138,20	20,73	41,46
,887	36,37	86,93	9,76	31,93
,518	19,68	51,28	5,70	15,54
,898	35,02	77,23	10,78	24,25
,771	28,53	59,37	7,71	29,30
1,151	54,10	111,65	20,72	46,04
1,094	32,82	103,93	14,22	33,91
,898	32,33	79,02	10,78	22,45
,910	10,92	60,97	3,64	21,84
1,133	29,46	86,11	11,33	30,59
1,140	22,80	99,18	9,12	22,80
1,082	46,53	108,20	17,31	37,87
,771	23,90	71,70	11,57	22,36
,956	38,24	87,95	13,38	30,59
1,140	35,34	88,92	7,98	34,20
1,094	39,38	109,40	13,13	43,76
1,059	22,24	73,07	6,35	32,83
1,036	45,58	68,38	19,68	42,48

Subsequently, a multivariate analysis of weighted variables will be performed using SPSS statistical software.

**Multivariate Tests<sup>a</sup>**

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	1,000	5951,330 <sup>b</sup>	1,000	1,000	,008
	Wilks' Lambda	,000	5951,330 <sup>b</sup>	1,000	1,000	,008
	Hotelling's Trace	5951,330	5951,330 <sup>b</sup>	1,000	1,000	,008
	Roy's Largest Root	5951,330	5951,330 <sup>b</sup>	1,000	1,000	,008
counselingweighted	Pillai's Trace	,999	44,224 <sup>b</sup>	18,000	1,000	,118
	Wilks' Lambda	,001	44,224 <sup>b</sup>	18,000	1,000	,118
	Hotelling's Trace	796,040	44,224 <sup>b</sup>	18,000	1,000	,118
	Roy's Largest Root	796,040	44,224 <sup>b</sup>	18,000	1,000	,118

a. Design: Intercept + counselingweighted

b. Exact statistic

**Tests of Between-Subjects Effects**

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	efficacyweighted	2888,685 <sup>a</sup>	18	160,482	44,224	,118
	qolweighted	8565,763 <sup>b</sup>	18	475,876	295,060	,046
Intercept	efficacyweighted	21596,294	1	21596,294	5951,330	,008
	qolweighted	149406,269	1	149406,269	92637,356	,002
counselingweighted	efficacyweighted	2888,685	18	160,482	44,224	,118
	qolweighted	8565,763	18	475,876	295,060	,046
Error	efficacyweighted	3,629	1	3,629		
	qolweighted	1,613	1	1,613		
Total	efficacyweighted	25054,132	20			
	qolweighted	160062,064	20			
Corrected Total	efficacyweighted	2892,313	19			
	qolweighted	8567,376	19			

a. R Squared = ,999 (Adjusted R Squared = ,976)

b. R Squared = 1,000 (Adjusted R Squared = ,996)

The effect of counseling on both high efficacy drug response and high quality of life has an overall trend to significance at  $p = 0.118$ , with  $p = 0.20$  taken as threshold. The separate p-values for the effects of counseling on respectively high efficacy drug response and high quality of life are 0.118 and 0.046. Slightly better statistics is provided if compliance is added to the analysis as a weighted least square factor for adjusting efficacy data with inconsistent spread (see also Chap. 25, Machine learning in medicine a complete overview, 2015, Springer Heidelberg Germany, from the same authors). The results tables of the weighted least squares procedure are given underneath.

**Multivariate Tests<sup>a,b</sup>**

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	1,000	6924,695 <sup>c</sup>	1,000	1,000	,008
	Wilks' Lambda	,000	6924,695 <sup>c</sup>	1,000	1,000	,008
	Hotelling's Trace	6924,695	6924,695 <sup>c</sup>	1,000	1,000	,008
	Roy's Largest Root	6924,695	6924,695 <sup>c</sup>	1,000	1,000	,008
counselingweighted	Pillai's Trace	,999	57,593 <sup>c</sup>	18,000	1,000	,103
	Wilks' Lambda	,001	57,593 <sup>c</sup>	18,000	1,000	,103
	Hotelling's Trace	1036,679	57,593 <sup>c</sup>	18,000	1,000	,103
	Roy's Largest Root	1036,679	57,593 <sup>c</sup>	18,000	1,000	,103

a. Design: Intercept + counselingweighted

b. Weighted Least Squares Regression - Weighted by complianceweighted

c. Exact statistic

**Tests of Between-Subjects Effects<sup>a</sup>**

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	efficacyweighted	87703,339 <sup>b</sup>	18	4872,408	57,593	,103
	qolweighted	273513,143 <sup>c</sup>	18	15195,175	404,126	,039
Intercept	efficacyweighted	585831,331	1	585831,331	6924,695	,008
	qolweighted	4052032,694	1	4052032,694	107766,434	,002
counselingweighted	efficacyweighted	87703,339	18	4872,408	57,593	,103
	qolweighted	273513,143	18	15195,175	404,126	,039
Error	efficacyweighted	84,600	1	84,600		
	qolweighted	37,600	1	37,600		
Total	efficacyweighted	858672,366	20			
	qolweighted	5157882,120	20			
Corrected Total	efficacyweighted	87787,939	19			
	qolweighted	273550,743	19			

a. Weighted Least Squares Regression - Weighted by complianceweighted

b. R Squared = ,999 (Adjusted R Squared = ,982)

c. R Squared = 1,000 (Adjusted R Squared = ,997)

In conclusion, the MANOVA demonstrated that the effect of counseling on the bivariate outcome drug efficacy and quality of life was significant at  $p = 0.118$  and

0.046. After weighted least square p-values fell to 0.103 and 0.039. Counseling simultaneously improved drug efficacy and quality of life.

## 19.5 Conclusions

In this chapter three examples are given of multivariate meta-analyses of therapeutic studies. In example 1 point estimates of the studies without measure of spread were meta-analyzed, in example 2 study samples sizes were included in the analysis as predictor variable, in example 3 study sample sizes were used for data weighting purposes instead of standard errors. In all of the examples a beneficial effect of the predictors on the multivariate outcomes was observed.

## Reference

More information and stepwise analyses of multivariate models is given in SPSS for starters and 2nd levelers second edition, Chap. 18, pp. 101–8, Springer Heidelberg Germany, 2016, from the same authors.

# **Chapter 20**

## **Transforming Odds Ratios into Correlation Coefficients**

### **Correlation Coefficients as a Replacement of Odds Ratios**

**Abstract** In order for odds ratios to be suitable for effect size estimation in meta-analysis, a measure of their spread is required. Unfortunately, this is often missing in clinical reports. A tentative solution for the problem is the replacement of odds ratios with regression coefficients and correlation coefficients, for which it is easier to compute a measure of spread. In this chapter the Yule and Ulrich approximations and the tetrachoric correlation coefficients are explained as possible solutions for odds ratios without measure of spread. This subject is pretty new, and possible solutions are, so far, little used in present meta-analyses.

#### **20.1 Introduction**

Odds ratios are often used to report the effect size of comparative therapeutic studies, and to perform meta-analyses of a multiplicity of such studies. The problem is, that a meta-analysis requires weighting of the effect sizes of the included studies, and, for that purpose, a measure of spread of the odds ratios has to be available. Unfortunately, in many studies the latter is missing. This chapter will review proposed solutions for this problem. All of these solutions are based on the fact, that a, commonly, applied, and, generally, accepted method for estimating the spread of correlation coefficients does exist. The standard error (se) of a correlation coefficient ( $r$ ) is given by:

$$\text{se of correlation coefficient } r = (1 - r^2) / \sqrt{n},$$

where  $r$  = correlation coefficient, and  $n$  = sample size of study. If we are able to transform unweighted odds ratios (ORs) into their best fit correlation coefficients, and, subsequently, calculate the standard errors of these correlation coefficients, then we will be able to perform meta-analyses on weighted correlation coefficients as outcome, instead of unweighted odds ratios. This chapter will review possibilities. All this is pretty new, but some studies have been published.

## 20.2 Unweighted Odds Ratios as Effect Size Calculators

This chapter is to show how odds ratios can be transformed into correlation coefficients or regression coefficients for weighting effect sizes of studies more easily. Odds ratios (ORs) are often used as measures of the effect size of comparative therapeutic studies. An OR of 1 means no difference between test treatment and control treatment. An OR >1 indicates that the test treatment is better than control, and, in addition, it gives an estimate about how much better. An OR of 2 would mean about twice times better, etc. However, if ORs are obtained from a very small data sample, two times better does not mean too much, and a test for statistical significance will be required, and, for that purpose, a measure of spread, e.g., the standard error (se) or the 95% confidence interval is needed. A common procedure is to make an estimate from the equation:

$$se_{\log OR} = \sqrt{(1/a + 1/b + 1/c + 1/d)},$$

where a to d are the number of patients per cell in a  $2 \times 2$  interaction matrix

	responder yes	no
test treatment	a	b
<u>control treatment</u>	c	d.

Unfortunately, in many published studies ORs are reported without information about the numbers of patients per cell, a to d. ORs without spread information are called unweighted ORs, and, those obtained from small studies, will disproportionately determine the final result of average ORs in a meta-analysis.

## 20.3 Regression Coefficients and Correlation Coefficients as Replacement of Odds Ratios

A possible solution to the problem of bias due to unweighted ORs is to replace them with correlation coefficients or regression coefficients. The effect size of comparative therapeutic studies, e.g., therapeutic drug trials, is, currently, usually measured as one of three estimators (1), (2), (3):

- (1) standardized mean difference, sometimes called Cohen's d or strict standardized mean difference (SSMD)
- (2) odds ratio,
- (3) correlation coefficient (otherwise called phi-values, if outcomes are binary).

Basic information of the use of the standardized mean difference (1) and the odds ratio (2), as effect sizes in a meta-analysis have already been reviewed in the Chap. 2. The correlation coefficient (3) is equal to the standardized regression coefficient. Its use in meta-analysis is pretty new and stems from the psychology literature. Although pretty explorative to clinicians searching for confirmative evidence, psychological meta-analyses have made valuable contributions to the mathematical approaches to meta-analysis, e.g., by including new effect size estimators to meta-analytic datasets (Peterson and Brown, *J Appl Psychol*, On the use of beta coefficients for meta-analysis, 2005; 90: 175–81, and Borenstein et al., Chap. 6, Effect sizes based on correlations, in: Introduction to meta-analysis Wiley & Sons, 2009).

In studies with correlation coefficients as main outcome, it is pretty easy to obtain an approximate standard error of the correlation coefficient with the help of the equation:

$$\text{se of correlation coefficient } r = (1 - r^2) / \sqrt{n},$$

where  $r$  = correlation coefficient and  $n$  = sample size of study. The correlation coefficients and their standard errors of the separate studies in a meta-analysis can, then, be further applied for a weighted effect size and heterogeneity tests with the help of the traditional inverse variance weight procedures (Chap. 2 of the current edition). A pretty elegant approach in the world of meta-analysis is the transformation of ORs into equivalent correlation coefficients, as proposed by several authors

(Digby, Approximating the tetrachoric correlation, *Biometrics* 1983; 39: 753–7,  
 Sanchez-Meca, Effect size indices for dichotomized outcomes in meta-analysis,  
*Psychol Methods* 2003; 8: 448–67,  
 Ulrich, On the correlation of a naturally and artificially dichotomized variable, *Br J  
 Math Stat Psychol* 2004; 57: 235–51),

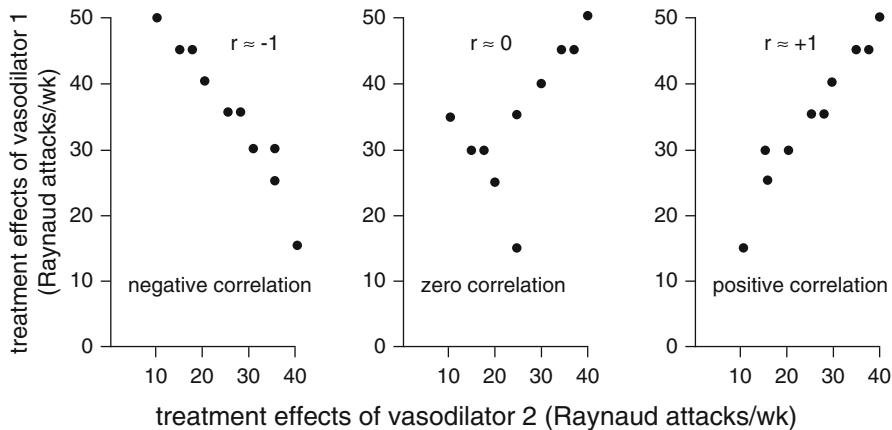
and summarized by:

Bonett (Transforming odds ratios into correlations for meta-analytic research, *Am Psychologist* 2007; 62: 254–55).

The correlation coefficient was, originally, named product-moment correlation coefficient by its inventor Pearson in the 1900 edition of the Philosophical Transactions of the Royal society of London, Series A, 195, pp. 1–147. It is the product of two mean-adjusted variables usually called  $x$  and  $y$ , and, it is thus, the covariance of the two variables, sometimes written as  $SP_{xy}$  (sum of products of  $x$  and  $y$  variables). The term moment refers to first moment about an origin, which is here the mean. In addition the correlation coefficient is standardized by dividing it by the square root of the product of the standard deviations (SDs) of  $x$  and  $y$ :

$$\text{correlation coefficient} = SP_{xy} / \sqrt{(SD_x^2 \cdot SD_y^2)}$$

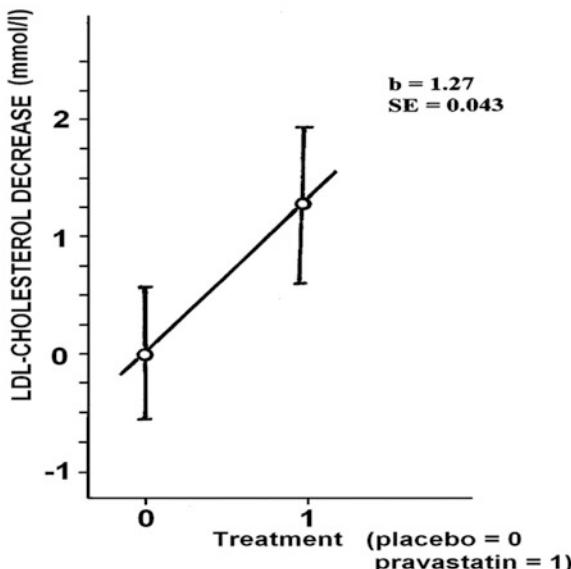
The above construct is pretty messy, but it leaves us with a quite wonderful functionality. It is, namely, a measure for the strength of linear association between the values from the x variable and the y variable. It varies from  $-1$  to  $+1$ . the strongest association is either  $-1$  or  $+1$ , the weakest association is zero.



The above graph gives an example of the results of three crossover studies comparing the effect of a test and control treatment in ten patients with Raynaud's phenomenon. Left the association is strong negative,  $r = -1$ , in the middle the association is zero,  $r = 0$ , right the association is strong positive,  $r = +1$ . Correlation coefficients can also be applied for the purpose of analyzing therapeutic treatment comparisons. An example is given underneath.

A parallel-group study of 872 patients studied for cholesterol decreasing capacity of pravastatin and placebo.

	placebo	pravastatin	difference
sample size	434	438	
mean (mmol/l)	-0.04	1.23	1.27
standard deviation (mmol/l)	0.59	0.68	standard error (se) = 0.043



The above graph shows, that the same result of the unpaired t-test for parallel-group analyses is obtained by drawing the best fit regression line for the study data.

The above table and the graph are from the data of the REGRESS study (Jukema, Circulation 1995; 91: 2528), a parallel-group study comparing the cholesterol reducing property of placebo versus pravastatin, the mean difference between the two treatments was 1.27 mmol/l with a standard error of 0.043 mol/l. This result can also be obtained by a linear regression with treatment modality, 0 or 1 for placebo or pravastatin, as x-variable and cholesterol reduction after treatment as outcome. However, the result of the regression analysis is expressed in the regression coefficient and its standard error,  $b$  and  $se_b$ -value, 1.27 and 0.043 mmol/l. Here the  $b$ -value seems to be useful as the effect size of the study, just like the mean difference between the two treatments is. Even more useful is the standardized  $b$ -value:

$$\begin{aligned} \text{standardized } b\text{-value} &= b\text{-value}/se_{b\text{-value}} \\ &= 1.27/0.043 \text{ standard error units.} \end{aligned}$$

The unit of the standardized  $b$ -value is often expressed in socalled standard error units. The standardized  $b$ -value is equal to the  $r$ -value of the linear regression from which it comes from.

$$\text{standardized } b\text{-value} = \text{correlation coefficient } r.$$

Correlation coefficients of individual studies can be used as the main outcome of comparative therapeutic studies. The meta-analysis of such studies requires information of the standard error of the correlation coefficients:

$$\text{standard error}_{\text{correlation coefficient } r} = (1 - r^2)/\sqrt{n}.$$

In this section we reviewed the use of correlation and regression coefficients for continuous outcomes. In the next section we shall see if they can also be used for binary outcomes.

## 20.4 Examples of Approximation Methods for Computing Correlation Coefficients from Odds Ratios

An example is given. In a hospital many patients tend to fall out of bed. We wish to find out, whether one department performs better than the other.

fall out of bed	yes	no
department 1	15(a)	20(b)
department 2	15(c)	5(d).

The ratio of the odds of falling out of bed in department 1 versus department 2 equals:

$$\text{OR} = 15/20 \div 15/5 = 0.25.$$

The correlation coefficient, here otherwise called phi because outcomes are binary (clinical data analysis on a pocket calculator second edition, Chap. 37, Phi tests for nominal data, 2016, Springer Heidelberg Germany, from the same authors),

$$\begin{aligned} &= (ad - bc) / \sqrt{(a + b)(c + d)(a + c)(b + d)} = \\ &= (75 - 300) / \sqrt{(35 \times 20 \times 30 \times 25)} = \\ &= -0.31 \end{aligned}$$

We can predict the chance of falling out of bed from the correlation coefficient. If it had been  $-1$ , the chance would have been 100%. If it had been  $0$ , the chance would have been 0%. Let us assume that in a study only the odds ratio (OR) of the departments is given:

$$\text{OR} = a/b \div c/d = 0.25.$$

### 20.4.1 The Yule Approximation

Yule (J Roy Stat Soc 1912; 75: 579) provided an equation for approximation of phi from OR:

$$\begin{aligned} \text{correlation coefficient} &= \\ (\text{OR}^{1/2} - 1)(\text{OR}^{1/2} + 1) &= (-0.5)(0.5 + 1) = -0.75 \end{aligned}$$

Obviously, this Yule approximation is a poor estimate, but this is due to the large difference in sample size of departments 1 and 2, and will otherwise perform better.

### 20.4.2 The Ulrich Approximation

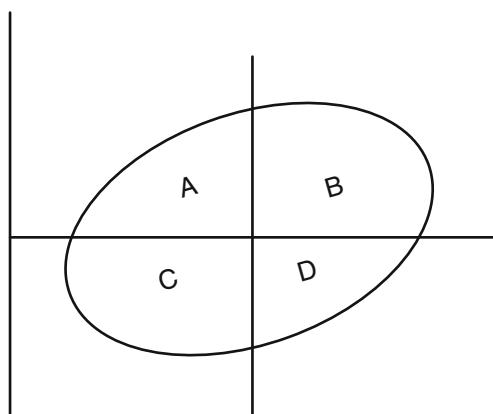
Suppose the samples sizes are known. Then, the Ulrich approximation (2004, Br J Math Stat Psychol) can be used, and it performs pretty well ( $-0.36$  as compared to  $-0.31$ ):

$$\text{correlation coefficient} = \log\text{OR}/\sqrt{(\log\text{OR}_2 + 2.89 n^2/n_1 n_2)} = -1.386/\sqrt{(1.92 + 12.49)} = -0.36$$

## 20.5 Computing Tetrachoric Correlation Coefficients from an Odds Ratio

Tetrachoric correlation coefficients are much similar to usual Pearson correlation coefficients, but they have another theoretical background, and provide better sensitivity of testing, if that background is sound and clinically supported.

If we have arguments to assume, that the binary data are from a continuous distribution, as is often the case in practice (e.g., negative means less severe, positive means above a threshold of severity), then a more appropriate approach will be to use a tetrachoric  $2 \times 2$  model, which is best described as an ellipse cut into four parts. The patients in the parts A + C are less severe according to the rater along the x-axis, those in the parts B + D are more severe. The patients in the parts A + B are more severe according to the y-axis rater, those in the C + D parts are less so. The vertical and horizontal lines are thresholds for respectively x- and y-axis raters. It is mathematically pretty hard to calculate the exact chance for patients of being in any of the parts A to D. It requires the concepts of eigenvectors and eigenvalues and bivariate normal distributions. And a lot of arithmetic, which is beyond the scope of the current work.



However, like with traditional correlation coefficients, approximation methods are possible. Like with a linear correlation between  $-1$  and  $+1$  as estimate for the strength of association of yes-no data according to two raters, a correlation coefficient can be calculated for the data in an ellipse-form pattern. It is called the tetrachoric correlation coefficient, ranging, just like the linear correlation coefficient, between  $-1$  and  $+1$ . But we should add a very pleasant aspect of the tetrachoric correlation assessments: they produce larger r-values, making the significance testing job more often successful.

Tetrachoric correlation assessments are used for the purpose of finding a correlation coefficient between two raters in a  $2 \times 2$  interaction matrix of the ellipse type. It was invented by Galton, a contemporary of Pearson, the inventor of the traditional correlation coefficient. The non-tetrachoric Pearson correlation coefficient, also usually called r or R, otherwise sometimes called Cohen's kappa, of the underneath  $2 \times 2$  table is commonly used for estimating the linear correlation between two strictly binary variables, and produces the following result. If an ellipse type interaction matrix is in the data, then the tetrachoric correlation coefficient should produce a larger correlation coefficients, because it better fits the data than does the traditional correlation coefficient.

		rater 1		
		diagnosis yes	no	
rater 2		40 A	10 B	50
		20 C	30 D	50
		60	40	100

$$\text{Pearson correlation coefficient } r = (70 - 50) / (100 - 50) = 0.4.$$

The tetrachoric correlation coefficient (tcc) is calculated much differently.

Approximation methods are usually applied, because exact computations are not available. The underneath approximation by Bonett (Transforming odds ratios into correlations for meta-analytic research, Am Psychologist 2007; 62: 254–55) makes use of the delta method, a mathematical approximation and logarithmic approximation where the variance of  $\log x$  is equal to the variance of x divided by x squared. This approach using the quadratic approximation and eigenvectors is sufficiently accurate, if samples are not too small.

$$\begin{aligned} \text{tcc} &= (\alpha - 1)/(\alpha + 1) \\ \alpha &= (AD/BC)^{\pi/4} = (1200/200)^{3.14/4} = 6^{0.785} = 4.08 \\ \text{tcc} &= (4.08 - 1)/(4.08 + 1) \\ &= 0.606 \end{aligned}$$

The tetrachoric correlation coefficient on a scale from 0 to 1 (or  $-1$ ) is, thus, much larger (0.606), than the traditional Pearson correlation coefficient r (0.400). The calculation by Pearson of the tetrachoric correlation coefficient given here is

based on a mathematical approximation. A more exact method is computationally also more complex, and Monte Carlo methods are preferred. Also a tetrachoric calculator is available on the internet (free tetrachoric calculator). Tetrachoric correlation is a special case of the polychoric correlation, applicable, when the observed variables are dichotomous. The tetrachoric correlation is also called the inferred Pearson Correlation from a two  $\times$  two table with the assumption of a bivariate normality. The polychoric correlation generalizes this to any larger table, the  $n \times m$  table, instead of a 2  $\times$  2 table.

## 20.6 Conclusion

Odds ratios are often used to report the effect size of comparative therapeutic studies, and to perform meta-analyses of a multiplicity of such studies. The problem is, that a meta-analysis requires weighting of the effect sizes of the included studies, and, for that purpose, a measure of spread of the odds ratios has to be available. Unfortunately, in many studies the latter is missing. This chapter reviews proposed solutions for this problem, based on the fact, that a, commonly, applied, and, generally, accepted method for estimating the spread of correlation coefficients does exist. If we are able to transform unweighted odds ratios into their best fit correlation coefficients, and, subsequently, calculate the standard errors of these correlation coefficients, then we will be able to perform meta-analyses on weighted correlation coefficients as outcome, instead of unweighted (and biased) odds ratios. This chapter will review possibilities.

It may be hard to transform an odds ratio into a correlation coefficient, unless information regarding its variance is known. Many studies, unfortunately, publish their results in the form of odds ratios without the additional information of variances. The variances are sometimes inferred from the 95% confidence intervals from graphs, however imprecise. Tetrachoric correlations can be demonstrated to be equal to b-squared values (= the best fit path statistics of dependent and independent variables squared). Usually Monte Carlo iteration of the 2  $\times$  2 table values leads to a useful b-value with standard error = unit (=1). In a meta-analysis, subsequently, the path statistics of the main results of the participating studies can, then, be used for finding the level statistical significance of the treatment efficacy versus zero. The tetrachoric correlation coefficients are much larger than the traditional Pearson correlation coefficients, a pleasant phenomenon.

The tetrachoric correlation coefficients can, then, be used for linear meta-regressions of meta-analyses of studies with odds ratios as outcome. All this is pretty new, but some studies have been completed and published (Mc Manus et al., Contrast-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies, BMC Med 2013; 11: 243. Williams, Using robust standard errors to combine multiple regression estimates with meta-analysis, 2012; [ecommons.luc.edu](http://ecommons.luc.edu)).

In this chapter three approximations are given for the calculation of correlation coefficients from studies where the measure of spread of their outcome, mostly odds ratios, are missing, the Yule, Ulrich and tetrachoric approximations.

## Reference

More information about the Yule, Ulrich and tetrachoric approximations of correlation coefficients are given in Yule (J Roy Stat Soc 1912; 75: 579), Ulrich (On the correlation of a naturally and artificially dichotomized variable, Br J Math Stat Psychol 2004; 57: 235–51), and Bonett (Transforming odds ratios into correlations for meta-analytic research, Am Psychologist 2007; 62: 254–55).

# Chapter 21

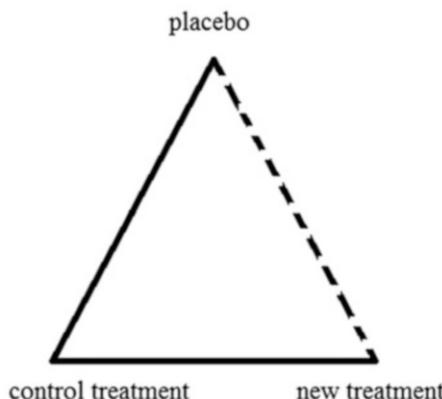
## Meta-analyses with Direct and Indirect Comparisons

### Challenging the Exchangeability Assumption

**Abstract** Meta-analyses indirectly comparing outcomes that have never been tested head to head is sometimes used as a solution for the lack of direct comparisons. In this chapter the Bucher method is explained for odds ratios and pooled odds ratios: the logodds ratio of control versus placebo is subtracted from the logodds ratio of new versus control, and this subtraction sum is adjusted for the pooled variances. This should give the best fit estimate of the comparison between the effect of new versus placebo. Our group recently proposed a slightly alternative method: the confidence intervals method for indirect comparisons. Real data examples of meta-analyses comparing new versus control, including noninferiority margins and previous trials of control versus placebo will be given.

### 21.1 Introduction

The high costs of trials, the continued introduction of new treatments, and ethical problems with placebo controls are causes for the increasing lack of direct comparisons in clinical research (Song et al, BMJ 2011; doi 101136), and meta-analyses indirectly comparing outcomes that have never been tested head to head, is sometimes used as a solution.



E.g., if a set of studies compared new treatment versus control treatment, and another set compared control treatment versus placebo treatment, then the pooled difference between new and control added to the pooled difference between control and placebo should provide information about a lacking direct comparison of new and placebo (dotted line in above graph). Three methods are available:

- (1) frequentists' methods.
- (2) network methods.
- (3) confidence intervals methods.

The network methods (2) include the Bayesian network methods already reviewed in the Chap. 12. But many more automatic data mining procedures can be used, like for example the automatic methods offered by the data mining work bench available in SPSS Modeler. In this chapter we will, particularly, review the frequentists' and confidence intervals methods.

## 21.2 Challenging the Exchangeability Assumption

We should emphasize that the methods of indirect comparisons are more relevant today than in the past, because of the availability of an increasing multitude of published studies, that can be compared with one another, either directly or indirectly. However, one assumption is the basis and bottleneck of all of this. That is the so-called exchangeability assumption, which means that studies to be compared should have exchangeable patient and study characteristics. If studies are not similar enough to be compared, all of these comparisons, and, particularly, the indirect ones that never took place in reality, will be pretty meaningless. The exchangeability assumption can sometimes be tested, but this is virtually always hard to do. E.g., a trial with three treatment arms A versus B versus C, should have the same effects of A versus C and B versus C as separate trials of two treatment arms should have. Currently, Extended Bucher networks (Song et al, BMJ 2011; 343: d4909) and Lumley networks (Stat Med 2002; 21: 2313–24) can cover for more complex networks with contrast coefficients adding up to one (see also the Chap. 22 of the current edition for a review of contrast coefficient meta-analysis).

## 21.3 Frequentists' Methods for Indirect Comparisons

Bucher proposed the adjusted indirect comparisons method for the purpose (J Clin Epidemiol 1997; 50: 683) for studies with odds ratios or pooled odds ratios as outcome of the indirect comparison of new treatment versus placebo:

$$\frac{\text{logodds ratio}_{\text{new versus control}} - \text{logodds ratio}_{\text{control versus placebo}}}{\text{Variancelogodds ratio}_{\text{new versus control}} + \text{Variancelogodds ratio}_{\text{control versus placebo}}}.$$

This approach requires, of course, assumptions:

- the separate comparisons are independent of one another,
- all of the trials measure the same effect,
- no differences must be in subgroup effects.

The above method has been successfully applied in a number of meta-analyses. We will name a few:

- Biondi et al. (Int J Cardiol 2011; 150: 325–31) included 32,893 patients in a meta-analysis of the indirect comparison of efficacy of antiplatelet therapies.
- Glenny et al. compared competing interventions in the health technology field (Health Technol Assess 2005; 9: 1–134.)
- Edward et al. reviewed randomised controlled trials for indirect comparisons using the same methodology (Int J Clin Pract 2009; 63: 841–54).

Instead of simple triangle indirect comparisons, Lumley (Network meta-analyses for indirect treatment comparisons, Stat Med 2002; 21: 2313–24) proposed and designed the analysis of multiple angle models, that are, however, essentially analyzed similarly to the triangle methods.

## 21.4 The Confidence Intervals Methods for Indirect Comparisons

A slightly different methodology for indirect comparisons was recently described by our group:

Cleophas and Zwinderman, A novel approach to non-inferiority testing: documented proof rather than arbitrary margins, Am J Ther 2012; 19: 101–7.

It uses confidence intervals as measure of spread instead of variances. It is, particularly, convenient for noninferiority assessments. Currently, new treatments are often tested against standard, because ethic committees do not allow for placebo controls, if a beneficial effect from a standard treatment has been unequivocally established. The problems of testing against active controls is, that differences are often small, and the new treatment may simply match the standard treatment in terms of efficacy. Investigators are aware of this phenomenon, and their attention is especially given to establishing ancillary advantages of the new compound, rather than extra efficacy benefits, although they have to make sure, that the efficacy of the new compound is, at least, equivalent, and definitely not inferior.

Kaul and Diamond (Ann Intern Med 2006; 145: 62) proposed, that placebo-controlled data of the standard treatment be added to any noninferiority trial, because a standard treatment without documented proof of superiority against placebo is not an adequate control treatment in a noninferiority study. In the current text we describe a pretty novel method, that includes a historical placebo-controlled study of the control treatment to be included in the analysis of the noninferiority study.

As an example, if a controlled trial of a standard treatment versus placebo produces a test statistic t- or z-value of 3.0 SEM-units, and an equally large trial with a new treatment versus a standard treatment produces a t- or z-value of say 0.5 SEM-units, then  $3 + 0.5 = 3.5$  SEM-units gives a good estimate of the comparison of the new treatment versus placebo. Or in summary:

$$\text{Standard versus placebo} = 3 \text{ SEM-units} \quad (1)$$

$$\text{New versus standard} = 0.5 \text{ SEM-units} \quad (2)$$

$$(1) + (2) = 3.5 \text{ SEM-units}$$

If the t-value of standard treatment versus placebo is very large, then even poorly performing new compounds may perform significantly better than placebo. If the t-value of standard treatment versus placebo is small, then the new compound will less easily outperform the placebo.

mean diff new vs stand (SEM-units)	mean diff stand vs placebo	mean diff new vs placebo
-2	+4	2 S
-1	+4	3 S
0	+4	4 S
+1	+4	5 S
+2	+4	6 S

mean diff new vs stand (SEM-units)	mean diff stand vs placebo	mean diff new vs placebo
-2	+2	0 NS
-1	+2	1 NS
0	+2	2 S
+1	+2	3 S
+2	+2	4 S

NS = not statistically significant; S = statistically significant; stand = standard treatment; new = new treatment; vs = versus; diff = difference.

The above table gives an overview of possible results of comparisons of the new treatment versus placebo using the confidence interval method. If the t-value of

placebo versus standard treatment is very large, then even poorly performing new compounds may be significantly better than placebo. If the t-value of placebo versus standard treatment is small, then the new compound will not easily perform better than placebo. If a noninferiority study is unable to demonstrate that, according to the above procedure, the new treatment is better than placebo, then the meaning of the noninferiority is very limited. The margin of noninferiority has probably been chosen too wide. We have to add here that the characteristics of the historical data should approximately match those of the new data, and that this information should be included in the report.

## 21.5 Real Data Examples

A real data example is given. A meta-analysis of two trials includes a noninferiority trial of new versus standard inhalers, and an earlier performed placebo controlled trial with the standard inhaler. The “New” and “Standard” used for the relief of asthma attacks compared in a non-inferiority study applied morning peak expiratory flow rates (l/min) as the primary measurement. The margin of noninferiority was set at  $-15$  l/min. The results of the trial were as follows:

Mean morning peak expiratory flow on treatment:

New = 420 ml/min (150 patients)

Standard = 416 ml/min (150 patients)

Mean difference between new and standard = 4 ml/min.

Estimated standard error of the mean of the difference, SEM = 5 ml/min.

1. The distance of the mean difference from the margin =  $-15 - 4 = -19$  ml/min =  $-19/5$  SEM-units =  $-3.8$  SEM-units. A t-value of  $-3.8$  SEMs-units corresponds to a p-value of 0.0001. Non-inferiority is demonstrated with a p-level as low as 0.0001.
2. A mean result of 4 ml/min =  $4/5 = 0.8$  SEM-units. The 95% confidence interval of this mean result  $0.8 \pm 2$  SEM-units is between  $-1.2$  and  $+2.8$  SEM-units, and does not cross the 0 value on the z-axis, and, so, the mean difference between standard and new is not significantly different from zero.
3. A similarly sized placebo-controlled trial of the standard treatment versus placebo produces a t-value of 3.0 SEM-units. The comparison of the new treatment versus placebo equals  $3.0 + 0.8 = 3.8$  SEM-units. The new treatment is, thus, significantly better than placebo at  $t = 3.8$  SEM-units, corresponding with a p – value of 0.0001. Both the lack of a significant difference between standard and new, and the significant difference between new and placebo support the presence of noninferiority of the new treatment versus the standard treatment.

## 21.6 Conclusion

The high costs of trials, the continued introduction of new treatments, and ethical problems with placebo controls are causes for the increasing lack of direct comparisons in clinical research, and meta-analyses indirectly comparing outcomes that have never been tested head to head, is sometimes used as a solution.

E.g., if a set of studies compared new treatment versus control treatment, and another set compared control treatment versus placebo treatment, then the pooled difference between new and control added to the pooled difference between control and placebo should provide information about a lacking direct comparison of new and placebo. The network methods include the Bayesian network methods already reviewed in the Chap. 12. In this chapter frequentists' and confidence intervals methods are reviewed.

We should emphasize that the methods of indirect comparisons are more relevant today than in the past, because of the availability of an increasing multitude of published studies, that can be compared with one another, either directly or indirectly. However, one assumption is the basis and bottleneck of all of this. That is the so-called exchangeability assumption, which means that studies to be compared should have exchangeable patient and study characteristics. If studies are not similar enough to be compared, all of these comparisons, and, particularly, the indirect ones that never took place in reality, will be pretty meaningless.

## Reference

More information of direct and indirect comparisons of clinical studies are given in the Chap. 12 of the current edition, entitled Network meta-analysis, and in the Chap. 63 in. Statistics applied to clinical studies 5th edition, Noninferiority testing, pp 675–86, 2012, Springer Heidelberg Germany, from the same authors.

# **Chapter 22**

## **Contrast Coefficients Meta-analysis**

### **Special Method for Assessing Random Effect Heterogeneity**

**Abstract** Contrast coefficients are arbitrary weights given to subgroups in a study, allowing for the identification of unexpected subgroup effects. Weights should add up to unit (1), and different contrast coefficient patterns can be tested for best fit of the data. An example of four studies of 10 patients per study assessing the fall in systolic blood pressure after different treatments was used. Meta-analyses including linear contrast testing provided better data fit than did traditional fixed and random effect meta-analysis.

#### **22.1 Introduction**

Traditionally, 95% confidence intervals of clinical studies are computed with mean  $\pm$  2 standard errors, under the assumption that outcomes have Gaussian frequency distributions. However, with notoriously heterogeneous studies like those of meta-analyses, this method lacks robustness (Bonett and Wright, Comments and recommendations regarding the hypothesis testing controversy, *J Org Behav* 2007; 28: 647–59). An improved model for the purpose with the help of linear contrast coefficients (Abdi and Williams, Contrast analysis, [www.utdallas.edu](http://www.utdallas.edu), 2010) has recently been proposed (Shuster, *Stat Med* 2010; 29: 1259–65, and Krizan, *Behav Res Meth* 2010; 42: 863–70). If your heterogeneity is assumed to be caused by a random effect of, e.g., 6 subgroups of 2 college and 4 noncollege students, then a simple contrast coefficient model will adjust the lack of robustness induced by the traditional model. With two studies of college students and 4 studies of noncollege students we have the following contrast coefficients respectively 1/2, 1/2, -1/4, -1/4, -1/4, -1/4 (should add up to zero). The confidence interval per study is now

$$\text{contrast coefficient } \times \text{mean} \\ \pm 2\sqrt{(\text{variance of the contrast coefficient } \times \text{mean})}$$

The null hypothesis of heterogeneity can be rejected, if ( $\Sigma$  of all studies in a meta-analysis = > 0).

$\Sigma$  contrast coefficient x mean

$\pm 2\sqrt{(\text{variance of the } \Sigma \text{ contrast coefficient x mean})}$

$= > 0.$

In this chapter, the use of contrast coefficients for testing heterogeneity in a meta-analysis will be assessed.

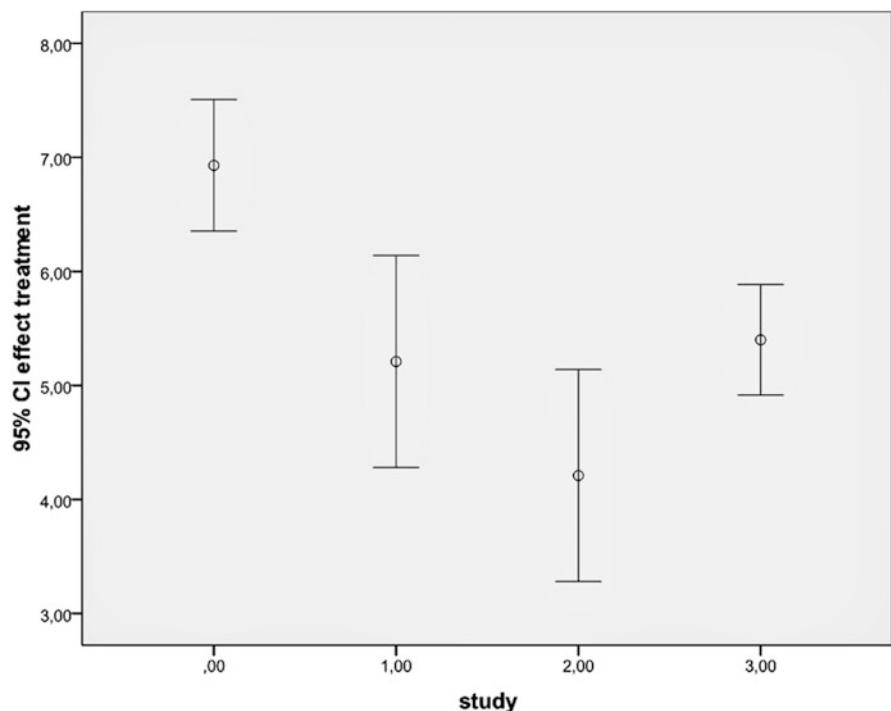
## 22.2 Example

As an example, in 4 studies involving 10 patients per study the fall in systolic blood pressure after treatment was the main outcome measure.

4 studies of 10 patients individual outcomes fall in syst blood pressure (mm Hg)

,00	6,00
,00	7,10
,00	8,10
,00	7,50
,00	6,40
,00	7,90
,00	6,80
,00	6,60
,00	7,30
,00	5,60
1,00	5,10
1,00	8,00
1,00	3,80
1,00	4,40
1,00	5,20
1,00	5,40
1,00	4,30
1,00	6,00
1,00	3,70
1,00	6,20
2,00	4,10
2,00	7,00
2,00	2,80
2,00	3,40
2,00	4,20
2,00	4,40
2,00	3,30
2,00	5,00
2,00	2,70
2,00	5,20
3,00	4,30
3,00	4,30

3,00	6,20
3,00	5,60
3,00	6,20
3,00	6,00
3,00	5,30
3,00	5,40
3,00	5,40
3,00	5,30



The above graph shows that the study means and 95% confidence intervals of four different studies assessing the fall in systolic blood after treatment are pretty heterogeneous.

### Report

effect treatment

study	Mean	N	Std. Deviation
,00	6,9300	10	,80561
1,00	5,2100	10	1,29910
2,00	4,2100	10	1,29910
3,00	5,4000	10	,67659
Total	5,4375	40	1,41615

The mean results are given above: study,00 = study 1, study 1,00 = study 2, etc.  
We will assess the studies for heterogeneity.

### 22.3 Fixed Effect Meta-analysis

The underneath table gives the computations for a fixed effect meta-analysis,  
var = variance.

	N	mean	standard error	1/var	mean/var	mean <sup>2</sup> /var
Study 1	10	6,93	0,2530	1.538	10.658	73.860
Study 2	10	5,21	0,4108	0.971	5.059	26.357
Study 3	10	4,21	0,4108	0.971	4.087	17.206
Study 4	10	5,40	0,2140	2.183	11.788	63.655
					+	
	40			5.663	31.59	211.078

$$\text{pooled mean} = 31.59 / 5.663 = 5.578$$

$$\text{Chi-square for pooled mean} = (31.59)^2 / 5.663 = 176.219$$

According to the chi-square table the p-value for 1 degree of freedom = < 0.0001.

Heterogeneity fixed effect model

$$\text{Pooled mean} = 5.578$$

$$\text{Chi-square value for pooled data} = 176.219$$

Heterogeneity of this meta-analysis is tested by the fixed effect model.

$$\begin{aligned} \text{Heterogeneity chi-square value} &= 211.078 - 176.219 \\ &= 34.859 \end{aligned}$$

With  $4 - 1 = 3$  degrees of freedom the p-value  
 $= < 0.0001$ .

## 22.4 Random Effect Meta-analysis

The underneath table gives the computations for a random effect meta-analysis (var = variance).

	N	mean	standard error	1/var	mean/var	mean <sup>2</sup> /var
Study 1	10	6,93	0.2530	1.538	10.658	73.860
Study 2	10	5,21	0.4108	0.971	5.059	26.357
Study 3	10	4,21	0.4108	0.971	4.087	17.206
Study 4	10	5,40	0.2140	2.183	11.788	63.655
						+
	40			5.663	31.59	211.078

$$\begin{aligned}\Sigma(\text{diff}^2/\text{var})/\text{N} - \Sigma(\text{diff}/\text{var})^2/\text{N}^2 &= 211.078/40 - (31.59/40)^2 \\ &= 5.277 - 0.624 \\ &= 4.653 \\ (\Sigma\text{diff}/\text{var})/(\text{N}) &= 31.59/40 \\ &= 0.790\end{aligned}$$

The variance of the random effect model  $s^*$  is given by

$$\begin{aligned}s^* &= 4 \text{ k/N} (1 + [(\Sigma\text{diff}/\text{var})/(\text{N})]^2)/8 \\ &= 4 \times 4/40 \times (1 + 0.79^2)/8 \\ &= 0.08 \\ \text{Heterogeneity chi-square value} &= 4 \times (4.653^2)/0.08 \\ &= 1082.5\end{aligned}$$

Now, with  $4 - 1$  degrees of freedom the above chi-square value produces a p-value of  $<0,0001$ . Both fixed and random effect heterogeneity tests are statistically very significant, and, thus, pooling these data make no sense. The results must be reported as a systematic review without weighted summary measures. If heterogeneity could be assumed to be due to random effect, then contrast coefficient analysis could have provided support for complex hypotheses like:

the studies are heterogeneous, because the data from the studies 2 and 3 are from academic hospitals, those of the studies 1 and 4 from community hospitals; study 1 included patients with mild hypertension only, the other three involved patients with more severe hypertension.

Instead of pocket calculator computations, we will perform the contrast coefficients analysis using one way analysis of variance in SPSS statistical software.

We will first perform a traditional one way analysis of variance. Start by entering your data.

**Now Command** click Menu....click Comparing Means....click One Way Analysis of Variance....Dependent List: enter effect treatment....Factor: enter study....click OK.

### ANOVA

effect treatment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	37,875	3	12,625	11,267	,000
Within Groups	40,339	36	1,121		
Total	78,214	39			

The above table shows that a very significant difference exists between the mean outcomes of the four studies, called groups by SPSS. We will now use contrast analysis to find out, whether a significant effect exists between the studies 2 and 3 (from academic hospitals), and the studies 1 and 4 (from community hospitals).

For that purpose we have to express a research hypothesis in the form of a contrast coefficient model.

## 22.5 Principles of Linear Contrast Testing

Traditionally, 95% confidence intervals are computed with mean  $\pm$  2 standard errors, under the assumption that outcomes have Gaussian frequency distributions. However, with notoriously heterogeneous studies like those of meta-analyses, this method lacks robustness (Bonett and Wright, Comments and recommendations regarding the hypothesis testing controversy, J Org Behav 2007; 28: 647–59).

Several improved models for the purpose with the help of linear contrast coefficients have been recently published:

- (1) Abdi and Williams, Contrast analysis, [www.utdallas.edu](http://www.utdallas.edu), 2010,
- (2) Shuster, Stat Med 2010; 29: 1259–65,
- (3) Krizan, Behav Res Meth 2010; 42: 863–70.

The 95% confidence intervals per study is computed as:

mean  $\pm$  95% confidence interval.

The null hypothesis of heterogeneity can be rejected:

$\Sigma$  mean  $\pm$  95% confidence interval  $>$  0.

Confidence interval is computed according to:

$\pm 2 \sqrt{(\text{variance of the mean})}$ .

If your heterogeneity is assumed to be caused by a random effect of, e.g., 6 subgroups of 2 college and 4 noncollege students, then a simple contrast coefficient model will adjust the lack of robustness induced by the traditional model.

With two studies of college students and 4 studies of noncollege students we have the following contrast coefficients respectively  $1/2$ ,  $1/2$ ,  $-1/4$ ,  $-1/4$ ,  $-1/4$ ,  $-1/4$  (should add up to zero). The confidence interval per study is now

$\text{contrast coefficient} \times \text{mean}$

$\pm 2\sqrt{\text{(variance of the contrast coefficient} \times \text{mean})}$ .

The null hypothesis of heterogeneity can be rejected if  $(\Sigma \text{ of all studies in meta-analysis})$

$\Sigma \text{ contrast coefficient} \times \text{mean}$

$\pm 2\sqrt{(\text{variance of the } \Sigma \text{ contrast coefficient} \times \text{mean})}$

is  $>0$ .

The above method performed better than the traditional method with simulated data. Bonett underscored the statement of Eysenck (An exercise in mega-silliness, Am Psychol 1978; 33: 1978) meta-analysis being pretty silly with careless selection and borderline quality of studies. This point is particularly relevant to the contrast coefficient method for calculating confidence intervals, which is rapidly meaningless under these circumstances.

Using this methodology, the statistical significance of a pooled results and statistical heterogeneity of multiple studies can be estimated with a t-table or chi-square table. Bonett gives some hypothesized examples (Meta-analytic interval estimation for standardized and unstandardized mean differences, Psychol Meth 2009; 14: 225–38).

## 22.6 Null Hypothesis Testing of Linear Contrasts

You first need a null-hypothesis. We will use the above example of four hypertension studies given in the introduction.

### Report

effect treatment

study	Mean	N	Std. Deviation
,00	6,9300	10	,80561
1,00	5,2100	10	1,29910
2,00	4,2100	10	1,29910
3,00	5,4000	10	,67659
Total	5,4375	40	1,41615

The above table gives the means and standard deviations of the studies 0, 1, 2, and 3.

The studies 0 and 3 were performed in the country, the other two in cities. The scientific question is, are the results of studies in the country different from those in the cities. The null-hypothesis to be tested will be:

The results of the country studies are not different from the city studies. The contrast is denoted as follows:

$$\text{Contrast} = -1 \text{ mean}_0 + 1 \text{ mean}_1 + 1 \text{ mean}_2 - 1 \text{ mean}_3$$

The null hypothesis:  $\text{Contrast} = 0$

$$\text{Contrast} = -1 \times 6930 + 1 \times 5210 + 1 \times 4210 - 1 \times 5400 = 0$$

One way analysis of variance can be adequately used for testing. We will first perform a pocket calculator analysis, and, then, contrast analysis as option of SPSS One Way Analysis of variance. SPSS requires, that the individual outcome data of the studies are available, which is not commonly the case.

## 22.7 Pocket Calculator One-Way Analysis of Variance (ANOVA) of Linear Contrasts

The underneath table give the pocket calculator one way anova for a linear contrast analysis of the data from the above 4 study meta-analysis. We will assess the hypothesis that a linear contrast exist between the studies [0 and 3] versus [1 and 2].

Study	mean	contrast coefficient (c)	mean times c	$c^2$
0 (n = 10)	6,93	-1	-6,93	1
1 (n = 10)	5,21	1	5,21	1
2 (n = 10)	4,21	1	4,21	1
3 (n = 10)	5,40	-1	-5,40	1 +
	0		-2,91	4

Computations are as follows.

$$\text{mean square}_{\text{between studies}} / \text{sum of squares}_{\text{within studies}} =$$

$$\text{mean square}_{\text{enumerator}} / \text{sum of squares}_{\text{denominator}}.$$

$$\begin{aligned} \text{sum of squares}_{\text{enumerator}} &= \frac{10(\sum (-1 \times 6,93 + 1 \times 5,21 + 1 \times 4,21 - 1 \times 5,40))^2}{\sum ((-1)^2 + (1)^2 + (1)^2 + (-1)^2)} \\ &= \frac{10(-2,91^2)}{4} = 84,7/4 \\ &= 21,4 \end{aligned}$$

For 1 degree of freedom:

$$\begin{aligned}
 \text{mean square enumerator} &= 21,4/1 = 21,4 \\
 \text{mean square denominator} &= \text{mean square error} \\
 \text{mean square error} &= \frac{(10-1)\text{SD}_{\text{study } 0}^2 + \dots + (10-1)\text{SD}_{\text{study } 3}^2}{\text{degrees of freedom} (=4n-4)} \\
 &= 9 (0,65 + 1,69 + 1,69 + 0,46)/36 \\
 &= 1125 \\
 \text{SD} &= \text{standard deviation} \\
 \text{F (Fisher) statistic} &= \text{mean square enumerator}/\text{mean square denominator} \\
 &= 21,4/1125 \\
 &= 19,02
 \end{aligned}$$

This F-statistic for linear contrast has 1 and 36 degrees of freedom, and is, thus, statistically very significant with  $p = 0.0001$  (see, e.g., Free p-Value Calculator for an F-Test at [www.danielsoper.com](http://www.danielsoper.com)).

## 22.8 Contrast Testing Using One Way Analysis of Variance on SPSS Statistical Software

The same data example is used. Open the SPSS statistical software program, and enter the data in the DATA View Screen.

### Then Command

click Menu....click Comparing Means....click One Way Analysis of Variance....Dependent List: enter effect treatment....Factor: enter study....click Contrast

Coefficients: enter -1....click Add....click 1....click Add....click 1....click Add....click -1....click Add....click continue....click OK.

The underneath table is in the output file, and shows the results of the contrast coefficients analysis, for which the SPSS one way anova menu does not use anova tests, but rather unpaired t-tests.

**Contrast Coefficients**

Contrast	study			
	,00	1,00	2,00	3,00
1	-1	1	1	-1

**Contrast Tests**

	Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
effect treatment	Assume equal variances	1	,66949	-4,347	36	,000
	Does not assume equal variances	1	,66949	-4,347	28,576	,000

The studies 1 and 4 have been given the contrast coefficients  $-1$  and  $-1$ . The studies 2 and 3 have been given the contrast coefficients  $1$  and  $1$ . Together they add up to zero. The test statistics for this contrast model equals  $-2,9100$ , and the t-value is much smaller than  $-1.96$ . This means that the outcomes of the studies 2 and 3 are largely different from those of the studies 1 and 4, and that this contrast explains the very significant heterogeneity of this meta-analysis. The t-value squared is approximately equal to the F-value, namely  $18,9$  and  $19,0$ . The pocket calculator, using SDs thus produced virtually the same results as the SPSS calculation using individual data of the separate studies did.

If the studies are not equal in size, a weighted contrast analysis will be required. The means of the studies 1 and 4 are estimated by  $(n_1 \text{ mean}_1 + n_4 \text{ mean}_4)/(n_1 + n_4)$  and those of the studies 2 and 3 by  $(n_2 \text{ mean}_2 + n_3 \text{ mean}_3)/(n_2 + n_3)$ . SPSS readily supplies the computations.

If you have arguments for contrasting group 0 versus the combined groups 2,3,4, then your model produces even better statistics. If you take into account less, e.g., equal variances, your test statistic further rises with a t-value of 6070.

**Contrast Coefficients**

Contrast	study			
	,00	1,00	2,00	3,00
1	3	-1	-1	-1

**Contrast Tests**

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
effect treatment	Assume equal variances	1	5,9700	1,15958	5,148	36	,000
	Does not assume equal variances	1	5,9700	,98357	6,070	21,045	,000

## 22.9 Conclusion

Contrast coefficients meta-analyses are particularly convenient for assessing random effect heterogeneity. As an example, in 4 studies involving 10 patients per study the fall in systolic blood pressure after treatment was the main outcome measure.

If heterogeneity could be assumed to be due to random effect, then contrast coefficient analysis could have provided support for complex hypotheses like:

the studies are heterogeneous, because the data from the studies 2 and 3 are from academic hospitals, those of the studies 1 and 4 from community hospitals; study 1 included patients with mild hypertension only, the other three involved patients with more severe hypertension.

Pocket calculator linear contrast testing with one-way analysis of variance is explained. It produced an F-value of 19.02 with 1 and 36 degrees of freedom ( $p = 0.0001$ ). Linear contrast testing with the help of SPSS statistical software produced the same magnitude of test statistic, a t-square value of 18.9 with 36 degrees of freedom.

## References<sup>1</sup>

- Abdi, Williams (2010) Contrast analysis. [www.utdallas.edu](http://www.utdallas.edu)  
Krizan (2010) Behav Res Meth 42:863–870  
Shuster (2010) Stat Med 29:1259–1265

---

<sup>1</sup>More information of contrast test models and novel models have been recently published.

# **Chapter 23**

## **Meta-analysis with Evolutionary Operations (EVOPs)**

### **Exploring the Effects of Small Changes in Experimental Settings**

**Abstract** Evolutionary operations (evops) tries and finds improved industrial processes by exploring the effect of small changes in an experimental setting. It stems from evolutionary algorithms, and uses rules based on biological evolution mechanisms.

In this chapter three subsequent studies of the determinants of infectious disease in eight operation rooms were studied. The effects of humidity, filter capacity change, and airvolume on numbers of infections was assessed. After the previous study, small changes in the experimental settings were made. The meta-data allowed for relevant conclusions about the optimization of infection free operation rooms.

#### **23.1 Introduction**

Evolutionary operations are sets of rules based on biological evolution mechanisms like mutation, recombination, and selection, they are used to analyze data files with thousands of variables, particularly genes. We live in a time of big data. Big data are so big, that they can no longer be handled by traditional computational computer programs. Every 20 months the entire database of the internet is doubled. From big data to meaningful information requires analytic methods. Traditional statistics applies averages. Machine learning applies teaching programs for computers. Computers are taught:

- to cluster data,
- to recognize patterns,
- to transfer (or not) data to a subsequent level.

Traditional meta-analyses come to the same conclusions, as those of the original studies. Meta-analyses of machine learning methodologies hardly exist, but this is a matter of time. Three important machine learning methodologies are:

- evolutionary operations is for improved industrial processes,
- Bayesian networks is for finding causes and consequences,
- support vector machines is the fast cluster methodology.

Bayesian networks and support vector machines have entered the world of meta-analysis. The two machine learning methodologies have already been reviewed in the current edition, and have respectively been used in network meta-analysis (Chap. 13), and in ensembled correlation coefficients and ensembled accuracies (Chaps. 18 and 19). This chapter will review the third machine learning methodology, evolutionary operations, which is virtually unused in medicine so far.

Evolutionary operations (evops) tries and finds improved industrial processes by exploring the effect of small changes in an experimental setting. It stems from evolutionary algorithms (Machine learning in medicine part three, Chap. 2, Evolutionary operations, 2013, and Machine learning in medicine a complete overview, Chap. 69, Evolutionary operations for process improvement, 2015, Springer Heidelberg Germany, both from the same authors), which uses rules based on biological evolution mechanisms. Each next generation is slightly different and generally somewhat improved as compared to its ancestors. It is widely used not only in genetic research, but also in chemical and technical processes. So much so that the internet nowadays offers free evop calculators suitable not only for the optimization of the above processes, but also for the optimization of your pet's food, your car costs, and many other daily life standard issues. In this chapter we will demonstrate that the ensembled analysis of evolutionary operation settings provides an optimization model for medical settings with multiple decision levels.

## 23.2 Example of the Meta-analysis of Three Studies Assessing Determinants of Infectious Disease

This chapter is to assess how evop methodology can be helpful to determine the optimal air quality of operation rooms. The air quality of operation rooms is important for infection prevention. Particularly, the factors humidity (1), filter capacity (2), and air volume change (3) are supposed to be important determinants. Can an evolutionary operation be used for process improvement.

Humidity, filter capacity and air volume change were linearly scored on 4 or 5 point scales: Humidity scores 1–5 (30–60%), (2) filter capacity scores 1–5 (70–90%), and (3) air volume change scores 1–5 (20–40% per hour). The protocol took care that all possible score combinations were evaluated separately. The first study only evaluated the scores 1 and 2, the second study the scores 3 and 4 for humidity and 2 and 3 for the rest, the third study the scores 4 and 5 for humidity and 3 and 4 for the rest.

### 23.3 First Study

Eight operation room air condition settings were investigated, and the results are underneath. We will use multiple linear regression with the number of infections as outcome and the three factors as predictors to identify the significant predictors.

humidity	filter capacity change	airvolume	infections
1,00	1,00	1,00	99,00
2,00	1,00	1,00	90,00
1,00	2,00	1,00	75,00
2,00	2,00	1,00	73,00
1,00	1,00	2,00	99,00
2,00	1,00	2,00	99,00
1,00	2,00	2,00	61,00
2,00	2,00	2,00	52,00

Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	145,500	15,512		9,380	,001
humidity1	-5,000	5,863	-,145	-,853	,442
filter capacity1	-31,500	5,863	-,910	-5,373	,006
airvolume change1	-6,500	5,863	-,188	-1,109	,330

a. Dependent Variable:infections1

The above table shows, that the humidity scores 1 and 2 and airvolume change scores 1 and 2 were not significant predictors of infections. Filter capacity scores 1 and 2 were, however, a significant predictor of infection at  $p = 0.006$ . The negative B value indicates, that the higher the filter capacity the fewer the numbers of infections.

### 23.4 Second Study

Again eight operation room air condition settings were investigated, but now slightly different scores were assessed. We will again use multiple linear regression with the number of infections as outcome and the three factors as predictors to identify the significant predictors.

humidity	filter capacity change	airvolume change	infections
3,00	2,00	2,00	51,00
4,00	2,00	2,00	45,00
3,00	3,00	2,00	33,00
4,00	3,00	2,00	26,00
3,00	2,00	3,00	73,00
4,00	2,00	3,00	60,00
3,00	3,00	3,00	54,00
4,00	3,00	3,00	31,00

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	103,250	18,243		5,660	,005
humidity2	-12,250	3,649	-,408	-3,357	,028
filter capacity2	-21,250	3,649	-,707	-5,824	,004
airvolume change2	15,750	3,649	,524	4,317	,012

a. Dependent Variable: infections2

The above tables shows that the humidity scores 3 and 4, airvolume change scores 2 and 3, and filter capacity scores 2 and 3 were all significant predictors of infections. The negative B values of humidity and filter capacity indicate, that the higher scores the fewer the numbers of infections. The positive B value of the airvolume change is unexpected, because you would expect that a higher airvolume would reduce rather than increase the numbers of infections. The effect could have been caused by a subgroup effect like confounding. When we performed a simple linear regression with infections as outcome and airvolume change as predictor, the p-value rose to 0.182, and, so, the significant effect was lost. Therefore, confounding, rather than a true effect, probably, caused the positive correlation here.

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	7,250	26,635		,272	,795
airvolume change2	15,750	10,447	,524	1,508	,182

a. Dependent Variable: infections2

## 23.5 Third Study

In the third study once again eight operation room air condition settings were investigated, and the scores were again slightly higher. Multiple linear regression is used with the numbers of infections as outcome and the three factors as predictors to identify the significant predictors.

humidity	filter capacity change	airvolume	infections
4,00	3,00	3,00	26,00
5,00	3,00	3,00	30,00
4,00	4,00	3,00	24,00
5,00	4,00	3,00	30,00
4,00	3,00	4,00	28,00
5,00	3,00	4,00	26,00
4,00	4,00	4,00	21,00
5,00	4,00	4,00	21,00

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	41,250	13,521		3,051	,038
humidity3	2,000	2,016	,299	,992	,377
filtercapacity3	-3,500	2,016	-,523	-1,736	,157
airvolumechange3	-3,500	2,016	-,523	-1,736	,157

a. Dependent Variable: infections3

The above tables shows that neither humidity scores 4 and 5, nor airvolume change scores 3 and 4, nor filter capacity scores 3 and 4 were significant predictors of infections anymore. Increasing the scores did, thus, not decrease the numbers of infections here anymore.

## 23.6 Meta-analysis of the Above Three Studies

The above results allowed for a series of conclusions:

- (1) increasing filter capacity from score 1 to 2 reduced the numbers of infections,
- (2) increasing humidity from score 3 to 4 reduced the numbers of infections,
- (3) increasing filter capacity from score 2 to 3 reduced the numbers of infections,
- (4) increasing airvolume change increased the numbers of infections,
- (5) increasing humidity score from 4 to 5 did not reduce the numbers of infections,
- (6) increasing filter capacity from score 3 to 4 did reduce the numbers of infections,
- (7) increasing airvolume change from score 3 to 4 didn't reduce infection numbers.

In summary, the humidity score 4 performed better than score 3. However, humidity score 5 did not perform better than score 4. Furthermore, filter capacity score 3 performed better than score 2. However, score 4 did not perform better than score 3. The scores for airvolume changes are hard to interpret probably due to confounding.

For now we will conclude that for optimization of infection free operation room settings it can be recommended to use a humidity score of 4 and a filter capacity score of 3, and that more information is needed regarding the recommended airvolume change settings.

In conclusion, evolutionary operations can be used to improve the process of air quality maintenance in operation rooms. This methodology can similarly be applied for finding the best settings for numerous clinical, and laboratory settings. We have to add, that interaction between the predictors was not taken into account in the current example, but that confounding could not be excluded. For a meaningful assessment of 2- and 3-factor interactions, larger samples would be required. Moreover, we have clinical arguments that no important interactions are to be expected. This pretty simple model already shows, that higher levels of humidity and filter capacity is not meaningful. Many more combinations of the four or five scores of 3 variables than the 24 ones applied in this example can be made. In order to assess the effect of airvolume changes, indeed, additional combinations have to be tested. However, testing is laborious, and, from the current meta-analysis we already know that airvolume change score from score 3 to 4 did not reduce the numbers of infections.

Meta-analyses of evops is quite different from the usual meta-analyses of interventional studies like drug efficacy studies. The latter, usually, comes to the same conclusion as those of the original studies, but with more power. With evops

things are very different, and meta-analysis is much more important than it is with interventional studies. For example, like in the example used, the first study may show, that score 3 is better than 2. The second study may show, that score 4 is not better than 3. From the combined analysis the conclusion would be, that the best option will be score 3.

## 23.7 Conclusion

Evolutionary operations are sets of rules based on biological evolution mechanisms like mutation, recombination, and selection, that are used to analyze data files with thousands of variables, particularly genes. We live in a time of big data. Big data are so big, that they can no longer be handled by traditional computational computer programs. Traditional statistics applies averages. Machine learning applies teaching programs for computers. Computers are taught things like data clustering, data pattern recognition, and data transfer to other levels.

Meta-analyses of machine learning methodologies hardly exist, but this is a matter of time. Evolutionary operations is a machine learning method for improved industrial processes, comparable with Bayesian networks for finding causalities and support vector machines the cleverest cluster method. Evolutionary operations (evops) works through exploring the effect of small changes in an experimental setting. Each next generation in an industrial process is slightly different and generally somewhat improved as compared to its ancestors.

Evolutionary operations is widely used not only in genetic research, but also in chemical and technical processes. So much so that the internet nowadays offers free evop calculators suitable not only for the optimization of the above processes, but also for the optimization of pet's food, car costs, and other daily life standard issues. In this chapter the ensembled analysis of evolutionary operation settings providing optimization models for medical settings with multiple decision levels is demonstrated.

## Reference

More background, theoretical and mathematical information of evops is given in Machine learning in medicine part three, Chap.2, Evolutionary operations, pp 11–18, Springer Heidelberg Germany, 2013, from the same authors.

# **Chapter 24**

## **Meta-analysis with Heterogeneity as Null-Hypothesis**

### **Automatic Data Mining in SPSS Modeler**

**Abstract** In traditional meta-analysis a single analysis-method is used for combined analysis of multiple studies with significant homogeneity of the studies as null-hypothesis. In this chapter a meta-analysis will be described for the combined analysis of open evaluation studies, with heterogeneity rather than homogeneity as null-hypothesis. An example of four studies receiving a different treatment per study, jointly including 120 patients with severe sepsis was used. Unlike traditional statistical methods, SPSS Modeler, a work bench for automatic data modeling, was able to identify the treatment 1 as the best of the four treatments with a predicted accuracy 91.8% of the decision tree output as applied.

#### **24.1 Introduction**

In the Chaps. 17 and 18 multiple analysis methods are applied for analyzing a single study with more power as main objective. In traditional meta-analysis a single analysis-method is used for combined analysis of multiple studies with significant homogeneity of the studies as null-hypothesis. In this chapter a meta-analysis will be described for the combined analysis of open evaluation studies, with heterogeneity rather than homogeneity as null-hypothesis. As an example four open evaluation will be used: the meta-analysis will include a combined analysis of the underneath types of data.

1. multiple continuous/binary/multinomial outcome data.
2. treatment modalities as predictor data.

Traditionally such types of data are analyzed with the underneath methods.

1. Linear regression with treatment modality as predictor and continuous variables as outcome. In the example of this chapter described underneath this method produced very significant benefits in three of the four treatment modalities.
2. Binary logistic regression with treatment as predictor and death as outcome, in the example of this chapter producing similar benefits.
3. Multinomial regression with treatments as predictor and low blood pressure as outcome, in the underneath example also similar benefits.

4. Multivariate analysis of variance with treatments as predictor and simultaneous analysis of multiple continuous, in the example also similar benefits.

The conclusions of the above analyses is that the three of the four treatments produced statistically significant benefits of survival, laboratory outcomes, and levels of blood pressures.

However, the best treatment is not immediately obvious, and subgroup analyses of the above multiple overall conclusions require at least a threefold number of additional statistical tests comparing the treatment 1 versus 2, 2 versus 3, and 1 versus 3. Bonferroni adjustment for multiple testing would mean that the null-hypothesis-rejection p-value would have to be lowered in order to maintain an overall type I error of 5%. After more than 5 tests the Bonferroni procedure may be appropriate but overconservative. Overconservative means that the new type I error (= the rejection p-value) will be artificially lowered, and consequently the type II error, otherwise called type II error, will rise, and, thus, the power (= 1-beta) will be lost, making the analysis pretty meaningless.

Using entropy based decision trees (Statistics applied to clinical studies, Chap. 53, Binary partitioning, pp. 579–586, Springer Heidelberg Germany, 2012, from the same authors) and Bayesian network based webs (Chap. 12), the best-fit cut-offs for best-fit estimators and their accuracies are computed. They do not provide p-values, but what they do is providing pretty-accurate quantitative information suitable for predictive purposes.

We should add, that these kinds of assessments are very much explorative in nature, and, that patient data of the individual studies are required. But, currently, we live in a world of big data, and, in many situations, obtaining patient data is no longer a problem.

## 24.2 Heterogeneity Rather Than Homogeneity of Studies as Null-Hypothesis

An example is given of a meta-analysis of four studies jointly including 120 patients with severe sepsis. One treatment per study was assessed. Multiple outcome variables were used to assess which one of the treatments performs best, but they were unable to identify the best treatment. As an alternative to the traditional statistical analysis methods as summarized in the above introduction, SPSS modeler, an analytics software application from IBM and work bench for automatic data mining and data modeling will be used.

The question “is a treatment a predictor of clinical improvement” is assessed by the question “is the outcome, clinical improvement, a predictor of the chance of having had a treatment”. This approach may seem incorrect, but is also used with discriminant analysis, and works fine, because it does not suffer from strong correlations between outcome variables (Machine Learning in Medicine Part One, Chap. 17, Discriminant analysis of supervised data, pp. 215–224, Springer

Heidelberg Germany, 2013). In this example, 120 patients with sepsis are treated with four different treatments. Various outcome values are used as predictors of the output treatment.

asat	alat	ureum	creat	crp	leucos	treat	low bp	death
5,00	29,00	2,40	79,00	18,00	16,00	1,00	1	0
10,00	30,00	2,10	94,00	15,00	15,00	1,00	1	0
8,00	31,00	2,30	79,00	16,00	14,00	1,00	1	0
6,00	16,00	2,70	80,00	17,00	19,00	1,00	1	0
6,00	16,00	2,20	84,00	18,00	20,00	1,00	1	0
5,00	13,00	2,10	78,00	17,00	21,00	1,00	1	0
10,00	16,00	3,10	85,00	20,00	18,00	1,00	1	0
8,00	28,00	8,00	68,00	15,00	18,00	1,00	1	0
7,00	27,00	7,80	74,00	16,00	17,00	1,00	1	0
6,00	26,00	8,40	69,00	18,00	16,00	1,00	1	0

asat = aspartate aminotransferase

alat = alanine aminotransferase

creat = creatinine

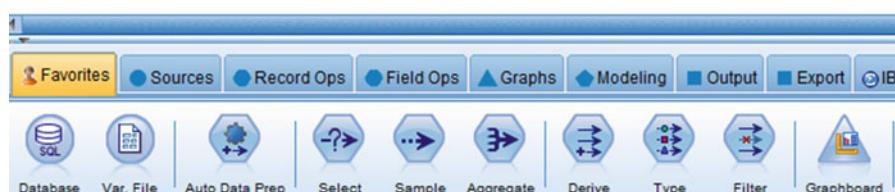
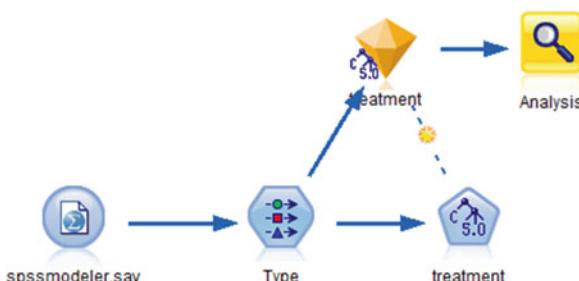
crp = c-reactive protein

treat = treatments 1–4

low bp = low blood pressure (1 no, 2 slight, 3 severe)

death = death (0 no, 1 yes)

Only the first 10 patients are above, the entire data file is in extras.springer.com and is entitled “spssmodeler3”. SPSS modeler version 14.2 is used for the analysis. We will start by opening SPSS modeler.



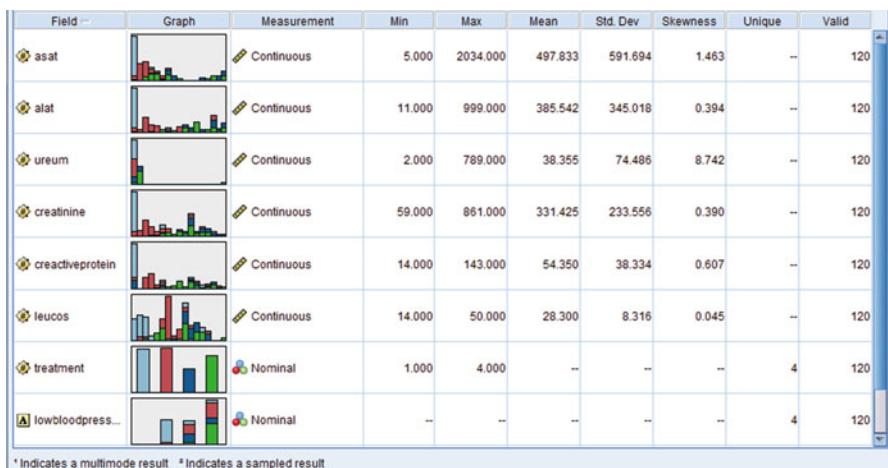
In the palettes at the bottom of the screen full of nodes, look and find the **Statistics File node**, and drag it to the canvas. Double-click on it...Import file: browse and enter the file “spssmodeler1.sav”...click OK...in the palette, find **Distribution node** and drag to canvas...right-click on the Statistics File node...a Connect symbol comes up...click on the Distribution node...an arrow is displayed...double-click on the Distribution Node...after a second or two the underneath graph with information from the Distribution node is observed.

### 24.3 Frequency Distribution of the Treatments



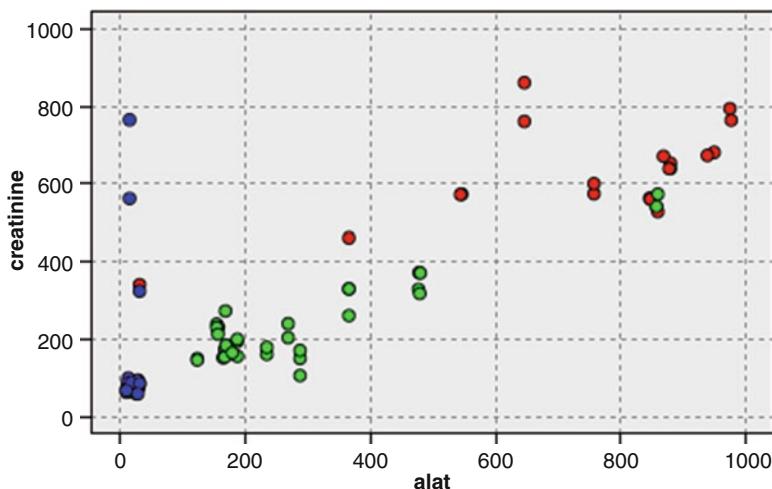
The Distribution node gives the frequency distribution of the four treatments in the 120 patient meta-data. All of the treatments are substantially present. Next we remove the Distribution node by clicking on it and press “delete” on the key board of your computer. Continue by dragging the Data audit node to the canvas.... perform the connecting manoeuvres as above...double-click it again.

### 24.4 Beneficial Effects of the Treatments, Histograms



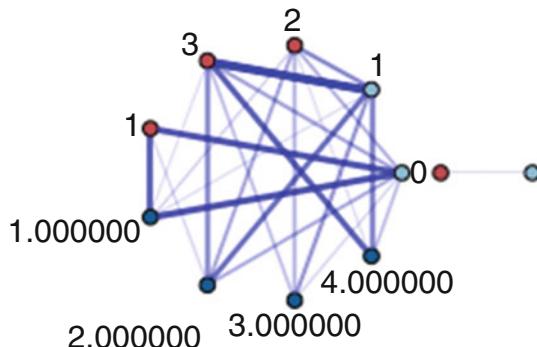
The Data audit node is helpful for demonstrating beneficial effects of the treatments. Click the Data audit node and select “treatment” as target field (field is variable here)....click Run. The information from this node is now given in the form of a Data audit plot, showing that, as a beneficial effect of the best treatments low values are frequently more often observed than high values. Particularly, the treatments 1 and 2 (light blue and red) are often associated with low values. These treatments are probably the best treatments. Next, remove the Data audit node by clicking on it, and press “delete” on the key board of your computer. Continue by dragging the Plot node to the canvas....perform the connecting manoeuvres as above....double-click it again.

## 24.5 Beneficial Effects of Treatments, Clusters



Also the Plot node is helpful for demonstrating beneficial effects of the treatments. Click the Plot node and in the Plot node tab select creatinine as y-variable and alat as x-variable, and treatment in the Overlay field at Color....click Run. The information from this node is now given in the form of a scatter plot of patients. This scatter plot of alat versus creatinine values shows that the four treatments are somewhat separately clustered. Treatment 1 (blue) in the left lower part, 2 (green) in the middle, and 3 in the right upper part. Low values means adequate effect of treatment. So treatment 1 (and also some patients with treatment 2) again perform pretty well. Next remove the Plot node by clicking on it and press delete on the key board of your computer. Continue by dragging the Web node to the canvas....perform the connecting manoeuvres as above....double-click it again.

## 24.6 Beneficial Effects of Treatments, Network of Causal Associations Displayed as a Web

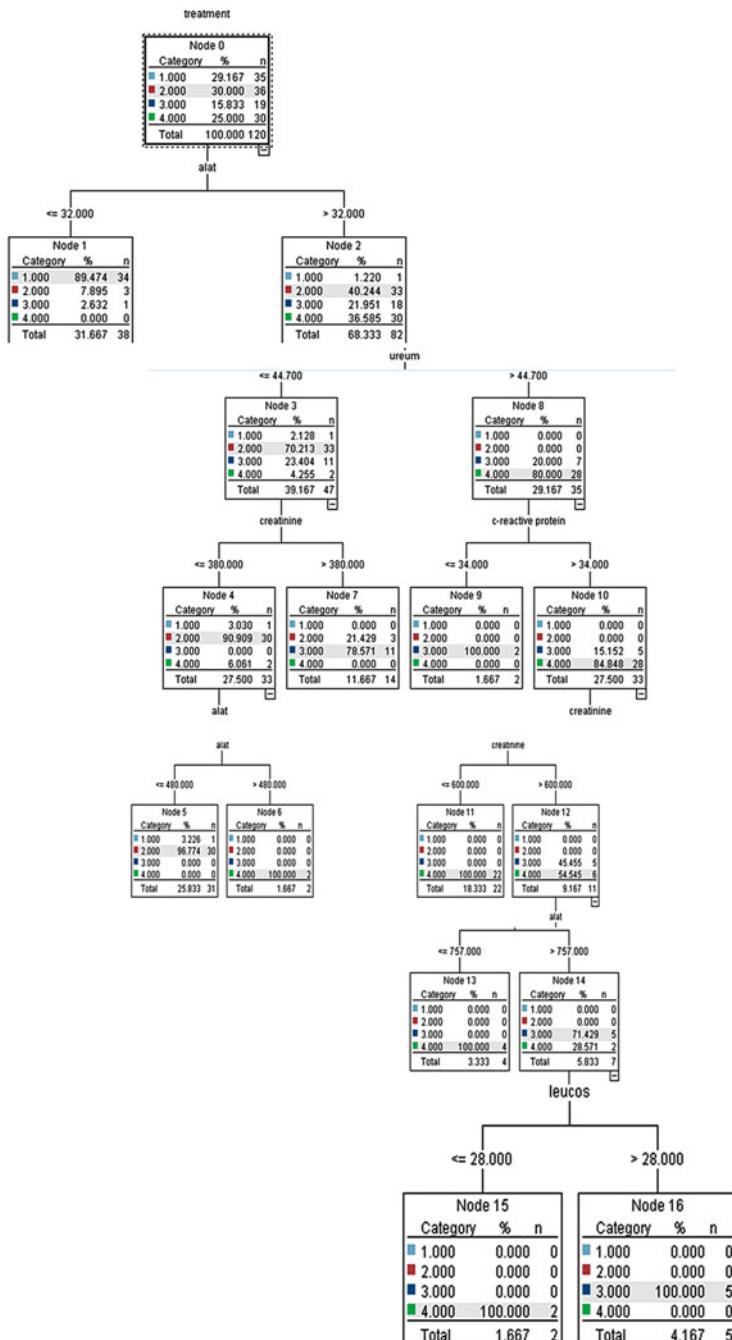


**Dark blue circles = treatment modalities (1.-4.000000).**  
**Red circles = increasing levels of low blood pressure (1-4).**  
**Light blue circles = death (1) and alive (0).**

Also the Web node is helpful for demonstrating beneficial effects of the treatments.

Just like Bayesian networks (Chap. 12), it is based on regression models with path statistics. Click the Web node and in the Web node tab click Select All...click Run. The web graph that comes up, shows that treatment 1 (indicated here as 1.000000) is strongly associated with no death and no low blood pressure (thick line), which is very good. However, the treatments 2 (2.000000) and 3 (3.000000) are strongly associated with death and treatment 2 (2.000000) is also associated with the severest form of low blood pressure. Next remove the Web node by clicking on it and press “delete” on the key board of your computer. Continue by dragging both the Type and C5.0 nodes to the canvas...perform the connecting manoeuvres respectively as indicated in the first graph of this chapter...double-click it again...a gold nugget is placed as shown above...click the gold nugget.

## 24.7 Beneficial Effects of Treatments, Decision Trees



The above output sheets give various interactive graphs and tables. One of them is the above C5.0 decision tree. C5.0 decision trees are an improved version of the traditional Quinlan decision trees with less, but more-relevant information. Decision trees use the entropy classification method whereby the parent node is repeatedly split into binary child nodes with best fit cut-off levels, i.e., those giving the smallest weighted impurity, otherwise called those with the fewest false positive and negative patients (see Machine learning in medicine a complete overview, Chap. 53, Decision trees for decision analysis, Springer Heidelberg Germany, 2015, from the same authors). In the current example the C5.0 classifier underscores the previous findings. The variable alat is the best classifier of the treatments with alat <32 in 89.5% of the patients having had treatment 1. The other classifiers performed less well, but high percentages of presence of the treatments 2–4 were generally associated with high laboratory scores, and, consequently, low presence of treatment 1 was observed.

## 24.8 Beneficial Effects of Treatments, Accuracy Assessment of Decision Tree Output

---

Results for output field treatment

Comparing \$C-treatment with treatment

Correct	112	91,8%
Wrong	10	8,2%
Total	122	

In order to assess the accuracy of the C5.0 classifier output an Output node is attached to the gold nugget. Find Output node and drag it to the canvas. . . . perform connecting manoeuvres with the gold nugget. . . . double-click the Output node again. . . . click Run. The output sheet shows an accuracy (true positives and true negatives) of 91,8%, which is pretty good.

## 24.9 Conclusions

Traditional statistical methods were unable to predict in a meta-analysis of four studies the best of four treatments. SPSS modeler can be adequately used for multiple outcomes analysis of clinical data. Finding the most appropriate treatment for a disease might be one of the goals of this kind of research. In the example of this chapter data mining methodologies including frequency distributions, histograms, Bayesian network based webs, entropy based decision trees were able to pretty accurately predict the best of four treatments for the treatment of severe sepses.

## Reference

SPSS modeler is a software program entirely distinct from SPSS statistical software, though it uses most if not all of the calculus methods of it. It is a standard software package particularly used by market analysts, but as shown can, perfectly, well be applied for exploratory purposes in medical meta-analysis and other medical research.

# **Chapter 25**

## **Meta-analytic Thinking and Other Spin-Offs of Meta-analysis**

### **Shift from P-Values to Effect Sizes of any Scientific Issue**

**Abstract** Terms like meta-learning and meta-analytic thinking are obviously spin-offs of meta-analysis methodology. At the same time meta-analysis has provided us with the ugliest graph in scientific research, the forest plot, otherwise called the blobbogram. Nonetheless, forest plots are currently increasingly applied for picturing patient characteristics in clinical studies, for propensity score assessments, for sensitivity assessments of the primary analysis in almost any clinical study. Pooled effect size assessments of important not only of the main outcomes of studies, but also of any other important or unimportant scientific issues. Another spin-off is, that the inclusion of a meta-analysis of the studies in the past are, currently, increasingly observed in novel interventional trials. This chapter will give examples of all of these spin-offs.

## **25.1 Introduction**

Modern meta-analyses do more than combine the effect sizes of a series of similar studies. The term “meta” in meta-analysis can be interpreted as “beyond”, and meta-analyses are currently increasingly applied for any analysis beyond the primary analyses of studies. In this chapter also terms like meta-learning, meta-analytic graphing, meta-analytic thinking, meta-cognition, meta-knowledge, meta-strategic knowledge, and awareness of learning processes will be addressed.

## **25.2 Meta-learning**

Kate (Cross-disciplinary perspective on meta-learning for algorithm selection, ACM Computing Surveys 2009; 41: doi 10.1145) gives some examples of novel methodologies and algorithms, including meta-cognition, meta-knowledge, higher order of thinking, meta-learning, meta-strategic knowledge, awareness of learning processes rather than knowledge and thinking skills. An important help to all of this is the methodology of forest plots. This month Nagendran et al. (Very large treatments effects in randomized trials, BMJ 2016; 355: i5432)

published a meta-analysis of 85,000 forest plots, and came to a conclusion never drawn before, ie., that controlled trials with very large effects were not good evidence givers of the progress of science. This suggests, that such trials, the constituents of the pinnacle of scientific research, may have been manipulated. So far, clinicians may have suspected something like that, and epidemiologists tended to comfort themselves by saying, that, in the end, there is something better than controlled trials, that is epidemiology with big data. Nowadays, we have stored big data, but storing is not enough. Particularly handling those big data with powerful methods like meta-analytic forest plots, Bayesian networks as already addressed in the Chap. 12, support vector machines, evolutionary operations technology as already addressed the Chap. 23, and automatic data mining programs already addressed in the Chaps. 17, 18, and 24, etc., can now provide a more rapid learning process of essential issues in the field of scientific progress.

Today every 20 s, a new scholarly article is published in the field of biomedicine (McGregor, director of Meta Science Search Machine, Toronto, 22-11-2016). Over the course of a year, that number, thus, will swell to more than 1.5 million. The pace of global research output has become too great to keep up with all of the medical products and tools that have become available to the scientific community. Now more than ever, new classes of tools, like the ones named above, are needed for confirmatory analyses and meta-analyses of novel compounds and other treatment modalities.

### 25.3 Meta-analytic Graphing

Forest plots are the most familiar graphing method for presenting the results of a meta-analysis, and can readily present in a single graph both the overall effect size and the effect size of the separate studies, together with their study sizes and confidence intervals. In addition, statistics like z-scores and p-values are often presented, and variations like summary forest plots, including cumulative forest plots, subgroup forest plots and “leave one out sensitivity” forest plots are increasingly added to traditional forest plots.

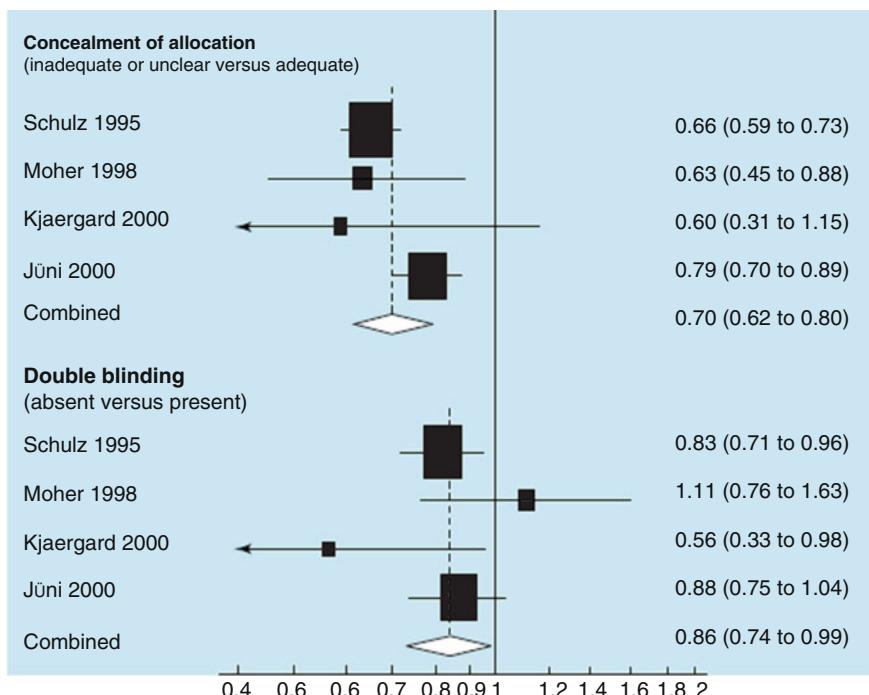
Forest plots are otherwise called blobbograms, a somewhat pejorative name stemming from the term blob = short cut for binary large object. However functional, it is by far the ugliest graph in the medical literature, and the name blobbogram would be consistent with the real meaning of the term blob being an amorphous mass. Maybe, the term blubbergram might even better cover the graphical method. A study from our group entitled “A meta-analysis of recent studies on patents admitted to hospital due to adverse drug effects (Int J Clin Pharmacol Ther 2009; 47: 549-56)” benefited from the blobbograms being replaced with tables.

The underneath graph is an example of an ugly blobbogram. It is taken from Juni et al., Empirical evidence of effect of study quality, BMJ 2001; 323: 42–6. It is a meta-analysis of 4 meta-analyses assessing the effects of blinding-adequacy on treatment efficacy in placebo-controlled studies. The study results were expressed as odds ratios. Nobody except gamblers understand the meaning of odds. The term stems from gambling, either you win or you lose, there is nothing in between. The real meaning of the ratio of two odds are even harder, but they are usually interpreted as the ratio of two chances: the chance of event with active treatment/chance of event with placebo. The problem with true chances is, that they run from zero to one, while odds run from zero to infinity. Statistical software programs have difficulties with chances and work fine with odds. The use of odds ratios has expanded through their use as major effect size estimators in meta-analyses.

The results of the inadequately and adequately blinded studies were pooled separately as odds ratios. The ratios of these pooled odds ratios were used for comparing the effect of inadequate blinding (“bad” studies) versus adequate blinding (“good studies”). For example, the underneath forest plot (copied from the public domain) consists of a meta-meta-analysis of 33 meta-analyses, 250 studies all in with odds ratios as study outcomes:

odds ratios = study outcomes  
ratios of odds = ratio of pooled odds ratios of “bad” studies and that of the ratios “good” studies.

The size of the black squares tells the magnitude of the separate meta-analyses, the straight lines on either side are the 95% confidence intervals of the pooled effects sizes of the studies, here expressed as odds ratios, the white diamond is the overall pooled result of the four meta-analyses with left and right angle again the 95% confidence intervals. The ratios were significantly lower than 1.0, on average 0.70 (95% confidence intervals 0.62–0.80). This would mean that inadequate blinding significantly decreased the hazard of eventful outcomes.



## 25.4 Meta-analytic Thinking in Writing Study Protocols and Reports

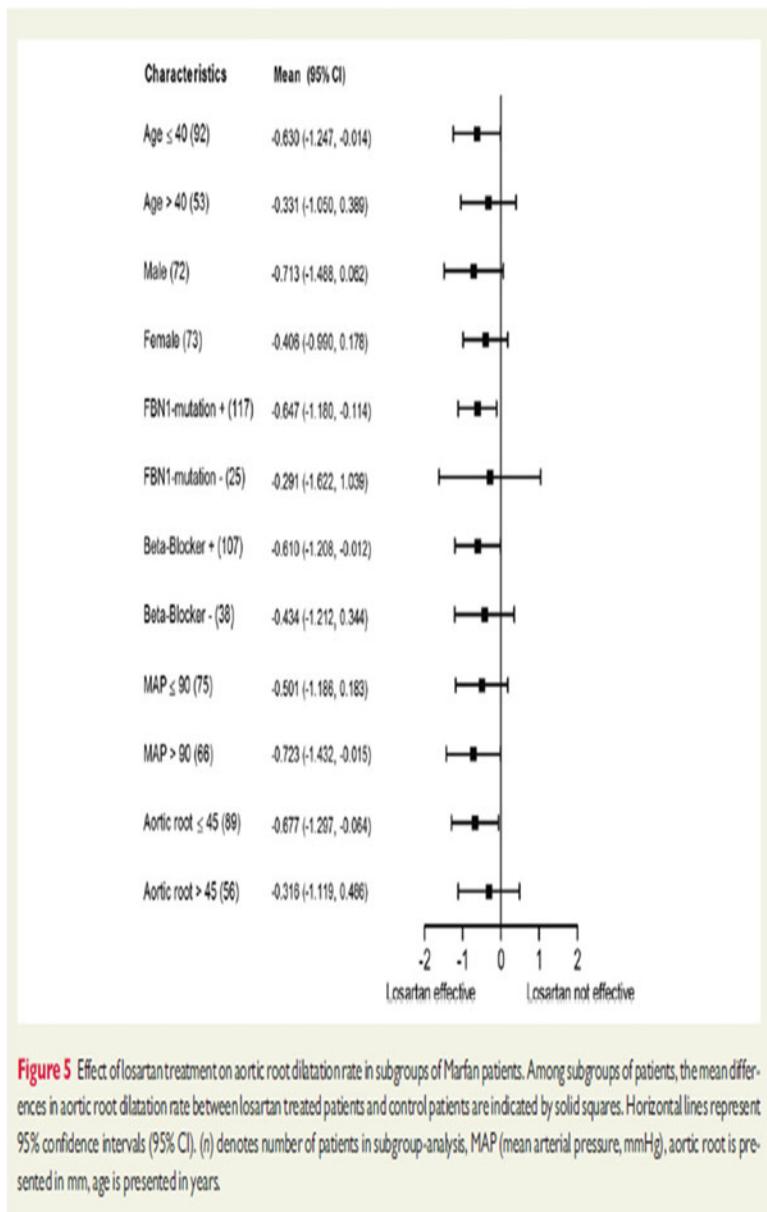
Trial protocols and reports in their methods or background sections may benefit from a meta-analysis of the literature to date. What must be in the introduction of your study protocol. First of all, we need background information. The background information is the main reason for the study. It must be in relationship with the current state of scientific knowledge. See the review of this process in Cleophas and Zwinderman, Understanding Clinical Data Analysis, Springer Heidelberg Germany, 2016, Chap. 2. What background information can be textually and conceptionally more complete and adequate than a meta-analysis of the peer-approved literature regarding the same subject as that of a study protocol. An example is given. The PhD thesis of Atiqi 2011, entitled Medicine a threat to health, University of Amsterdam Netherlands, started with background information in the form of the underneath meta-analysis of studies assessing the percentage of hospital admissions due to adverse drug effects. Then, a novel studies of a population of 2000 patients admitted through the emergency room was performed. Of the

entire population of 2000, 380 were definitely due to adverse drug effects (19%, with a 95% confidence interval 17–22%).

Table of studies on patients admitted to hospital due to adverse drug effect

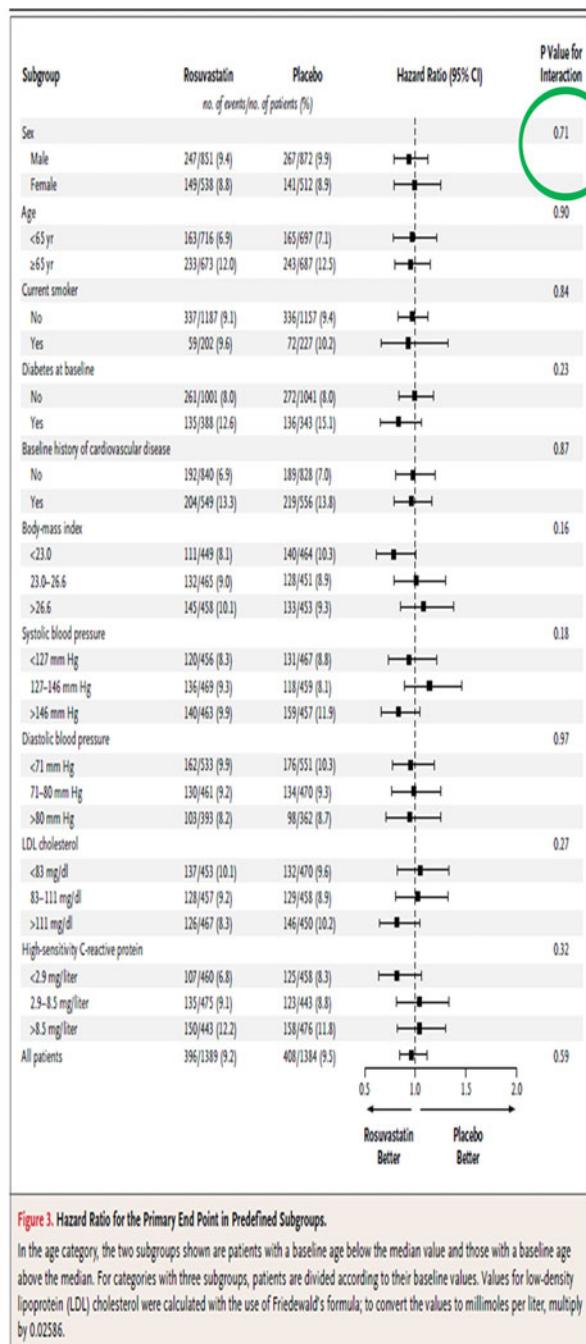
Study	study size	percentage of all admissions	95% confidence intervals
1.Mannesse et al 2000	106	21.0 %	13.0-29.0 %
2. Malhotra et al 2001	578	14.4 %	11.5-17.3 %
3. Chan et al 2001	240	30.4 %	24.5-36.3 %
4. Olivier et al 2002	671	6.1 %	4.3-7.9 %
5. Mjorndal et al 2002	681	12.0 %	9.5-14.5 %
6. Onder et al. 2002	28411	3.4 %	3.2-3.6 %
7. Koh et al 2003	347	6.6 %	4.0-9.2 %
8. Easton-Carter et al 2003	8601	3.3 %	2.9-3.7 %
9. Dormann et al 2003	915	4.9 %	4.9-14.3 %
10.Peyriere et al 2003	156	9.6 %	4.9-14.3 %
11.Howard et al	4093	6.5 %	5.7-7.3 %
12. Pirmohamed et al	18820	6.5 %	6.2-6.8 %
13.Hardmeier et al 2004	6383	4.1 %	3.6-4.6 %
14.Easton et al 2004	2933	4.3 %	3.6-5.0 %
15.Capuano et al. 2004	480	3.5 %	1.9-5.1 %
16.Caamano et al 2005	19070	4.3 %	3.7-4.6 %
17.Yee et al 2005	2169	12.6 %	11.2-14.0 %
18.Baena et al 2006	2261	33.2 %	31.2-35.2 %
19.Leendertse et al 2006	12793	5.6 %	5.2-6.0 %
20.Van der Hooft et al 2008	355	5.1 %	2.8-7.4 %
Pooled	113203	5.4 %	5.0-5.8 %

## 25.5 Meta-analytic Forest Plots of Baseline Patient Characteristics



**Figure 5** Effect of losartan treatment on aortic root dilatation rate in subgroups of Marfan patients. Among subgroups of patients, the mean differences in aortic root dilatation rate between losartan treated patients and control patients are indicated by solid squares. Horizontal lines represent 95% confidence intervals (95% CI). (n) denotes number of patients in subgroup-analysis, MAP (mean arterial pressure, mmHg), aortic root is presented in mm, age is presented in years.

The above forest plot is an example of a study of the effect of losartan versus placebo on aortic root dilatation rate (Groenink et al, Eur Heart J 2013; 34: 3491–500). The difference in dilatation rate after treatment was measured in the entire study population, and in separate patient groups with the above characteristics. With all of the characteristics the standardized mean difference was negative (meaning that losartan tended to perform better than placebo). A significant difference, as shown by a 95% confidence interval not crossing the zero point on the x-axis, was sometimes observed. This suggests the presence of interaction, otherwise called synergism, between some characteristics and the treatment modalities: with some of the characteristics the active treatment performed better than it did with others. The blobbogram is a very helpful device here.

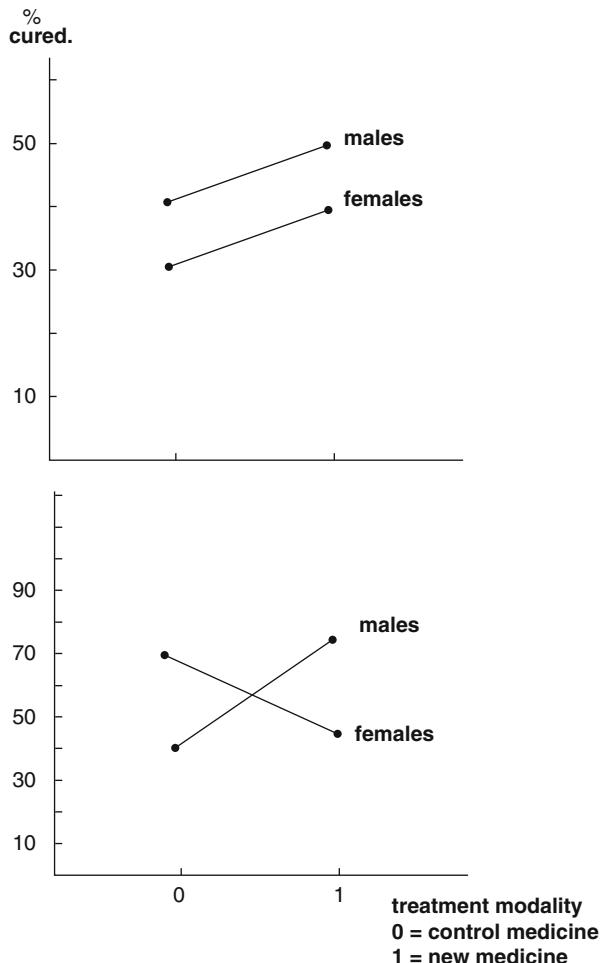


The above forest plot is another example of a study of the effect of rosuvastatin versus placebo on cardiovascular events in hemodialysis patients (Fellstrom et al, N Engl J Med 2009; 360: 1395–1407). The difference in the chances of cardiovascular events after treatment was measured in the entire population and in the subgroups with the above characteristics. With all of the characteristics the chance of events while on active treatment and placebo treatment were measured as hazard ratios. Hazard ratios were frequently smaller than 1, indicating less events while on rosuvastatin. This would be consistent again with interaction between some subgroup characteristics and the rosuvastatin treatment modality, but with none of the characteristics a statistical significance was obtained.

The above examples show that blobbograms invented for the graphical representation of the pooled results of a meta-analysis of multiple studies can very well be applied for the purpose of displaying subgroups different efficacies in a controlled trial to one of the treatments, otherwise called interaction or synergism between subgroups and treatment modalities. A significant difference, as shown by a 95% confidence interval not crossing the zero point on the x-axis, was sometimes observed. This suggests the presence of interaction, otherwise called synergism, between some subgroups and the treatment modalities in this study: in some of the subgroups the active treatment performed better than it did in other subgroups. The blobbogram is a very helpful device here.

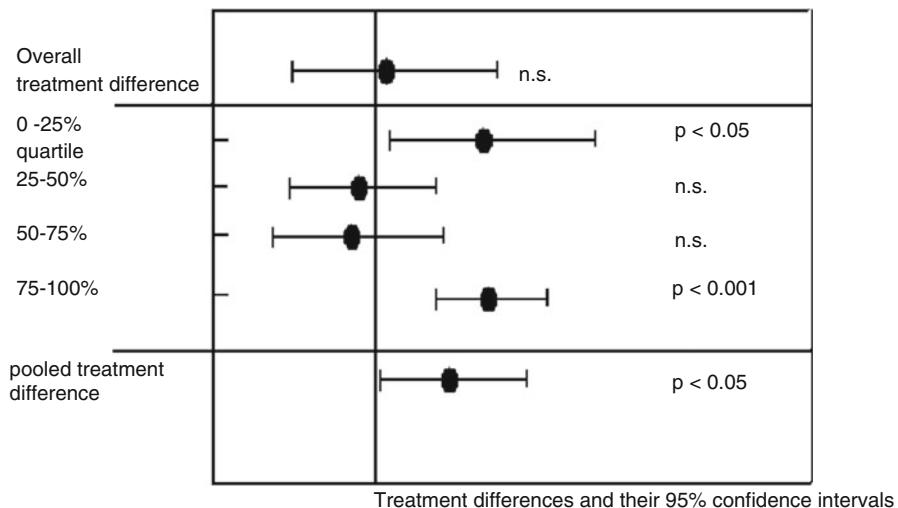
## 25.6 Meta-analysis of Forest Plots of Propensity Scores

A forest plot can be used not only for demonstrating interaction between subgroups and treatment modalities of controlled clinical trials, but also for demonstrating and managing another important phenomenon, i.e., confounding.



Confounding and interaction are different things. The above upper and lower graphs give respective examples of confounding and interaction. Confounding means that a subgroup performs better to every treatment given. This has a peculiar effect on the data analysis, if many males received the control, and many females the new treatment. It causes an overall line to be almost horizontal, and a treat efficacy to be partly or even entirely obscured.

Interaction means that one subgroup performs better with one treatment, the other subgroup does so with the other treatment. Again, an overall line can and often will become almost horizontal, obscuring the treatment efficacy. Blobbograms can not only be, obviously, applied for studying interaction, but also for visualizing and managing confounding.



The above forest plot gives an example of a propensity score procedure for adjusting multiple confounders. In a study with a single confounder, subclassification is easy. You, e.g., calculate the treatment differences between the males and the females, and then calculate a weighted mean in the same way as a pooled result is obtained in a traditional meta-analysis. If you have few confounders, multiple linear regression with the treatment modalities and the confounders as predictors is possible. However, if you have more than just a few, e.g., more than 3 or 4 confounders, then multiple regression becomes powerless, and propensity scores must be applied for adjusting purposes.

Propensity scores methodology requires, that, for every subgroup, the chance of having had treatment 1 versus that of treatment 2, or, rather, their odds (used as relative chances) must be calculated.

	treat 1 n = 100	treat 2 n = 100	chance treat 1 / chance treat 2 or their odds ratio (OR)	p (vs OR = 1)
1.Age>65	63 (%)	76	0.54 (63/37 / 76/24)	0.05
2.Age<65	37	24	1.85 (1/OR1 )	0.05
3.Dm	20	33	0.51	0.10
4.No DM	80	67	1.96	0.10
5.Smoker	50	80	0.25	0.10
6.No smoker	50	20	4.00	0.10
7.Hypertension	60	65	0.81	ns
8.No hypertension	40	35	1.23	ns
9.Cholesterol	75	78	0.85	ns
10.No cholesterol	25	22	1.18	ns
11.Renal failure	12	14	0.84	ns
12.No renal failure	88	86	1.31	ns

OR = odds ratio, DM = diabetes mellitus, treat = treatment, ns = not statistically significant. Then multiply the ratios of the odds (ORs) for each patient, that is, if they are statistically significant (see Chap. 3, Statistics applied to clinical studies 5th edition, Springer Heidelberg Germany, 2012, from the same authors).

Patient	old y/n	dm y/n	smoker y/n	prop score = OR1 x OR2 x OR...
1	y	y	n	$0.54 \times 0.51 \times 4 = 1.10$
2	n	n	n	$1.85 \times 1.96 \times 4 = 14.5$
3	y	n	n	$0.54 \times 1.96 \times 4 = 3.14$
4	y	y	y	$0.54 \times 0.51 \times .025 = 0.06885$
5	n	n	y	...
6	y	y	y	...
7	....			
8	....			

OR = odds ratio

y = yes

n = no

prop = propensity

Multiplication terms are called propensity scores. They tend to have largely different sizes. The next thing you do is, classifying the patients into 4 or more groups according to the size of their multiplication terms, otherwise called the propensity scores. Then calculate the treatment result per group, and calculate an overall treatment result, weighted. Weighting is similar to the procedure with one confounder. The above graph shows that the overall weighted treatment result is larger than the unadjusted treatment result. The overall treatment result gives a better picture of the real treatment effect, because it is adjusted for confounding.

## 25.7 Meta-analytic Thinking: Effect Size Assessments of Important Scientific Issues Other Than the Main Study Outcomes

Meta-analyses are currently often performed for purposes other than increasing the power of rejecting the null hypotheses of the primary studies, e.g., for assessing the effect sizes of all kinds of relevant scientific issues of studies other than their main outcome, and more quantitative information about biases of clinical research in general is, thus, obtained.

### 1. Effect size of placebo effects on the studies' outcome.

For example, Kirsch et al. meta-analyzed the magnitude of the placebo effect in 26 placebo controlled studies of antidepressants on improvements of depression scores after treatment. The placebo effect measured as mean depression scores on placebo versus active treatment up to 80% were observed (Kirsch et al, PLOS Med 2008; Doi: 10137).

### 2. Effect size of blinding.

The example of Sect. 25.2 of this chapter shows how a ratio of odds ratios can be used for the purpose.

### 3. Effect size of randomization.

The Chap. 6 example shows how in the study of Masoor and Cleophas (J Card Fail 2009; 15: 305–9) the pooled odds (and risk) ratios of diabetics versus non-diabetics in 10 observational and 7 randomized controlled clinical trials are respectively 1.22 and 1.35, with a pooled odds ratio of 1.28, all of them at  $p < 0.001$  different from an odds ratio of 1.0.

### 4. Effect size of the prospective nature in prospective meta-analyses.

The Chap. 10 reviews the meta-analysis of 22 case-control studies and 11 cohort studies of the risk of coronary artery disease in patients with and those without homocysteinemia. The studies were, of course, respectively retrospective and prospective in design. The pooled odds (and risk) ratios were respectively 1.62 and 1.49, with 1.58 overall, all of the values at  $p < 0.001$  versus an odds ratio of 1.0.

### 5. Effect size of “outcome reporting bias”.

Publication bias means that negative trials are at risk of not being published or being published (much) later than their positive counterparts. Publication bias assessments is a standard procedure in any meta-analysis (see Chaps. 1 and 2 for examples). In the past few years another closely related type of bias, has been increasingly recognized, the “bias due to selective reporting”, otherwise called “outcome reporting bias”. Kirkham et al. (BMJ 2010; 340: c365. doi:1136) published a meta-analysis. The studies included in 33 systematic reviews from the Cochrane Library were classified based on insignificant results, and lack of reporting their results. They were classified as high, low, and no risk of reporting bias. A strong effect of reporting bias was established with very significant differences between pooled outcome estimates with versus those without suspected reporting bias.

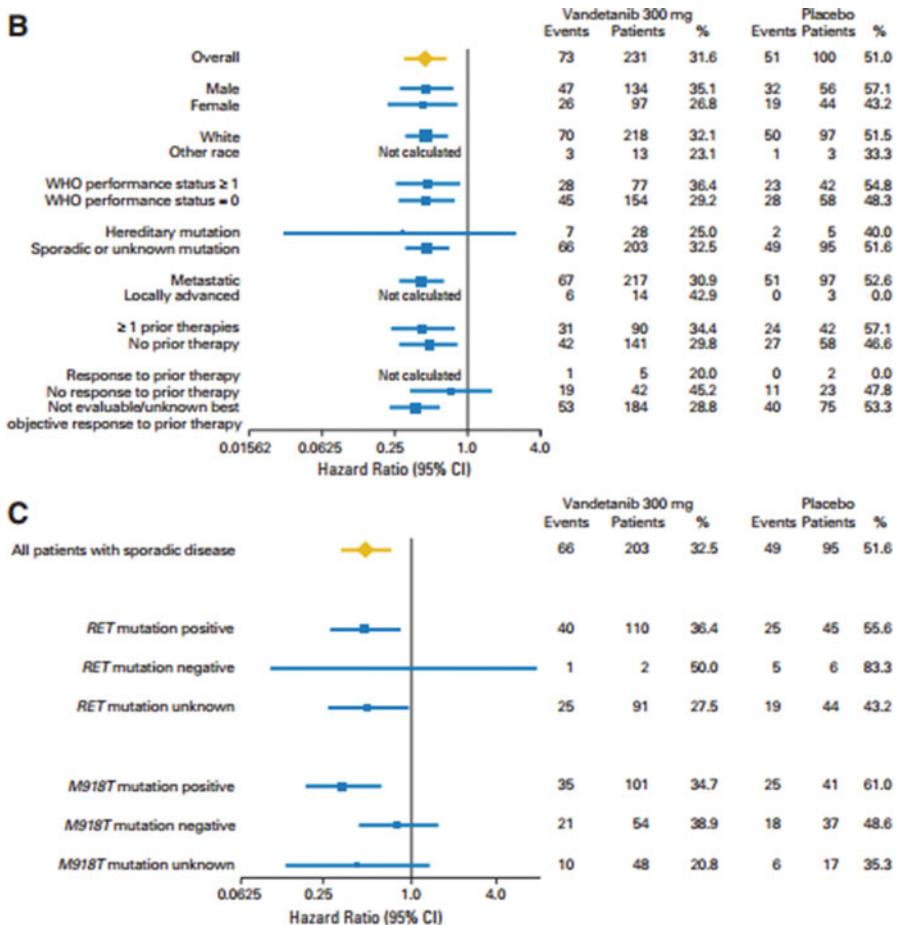
### 6. Effect sizes of overtly optimistic abstracts, conclusion based on subgroups, discrepancies between aims and conclusions of studies etc.

The phenomenon called spin, indicating specific reporting strategies either intentionally or not, to convince readers that the benefits of a trial are greater

than shown by the results, is closely related to that of reporting bias. Lazarus et al (2015, Doi: 11.1186, BMC Med Res Methodol) reviewed 126 studies for the purpose. Not all of the above mentioned effect sizes although relevant meta-analytic quantities have been widely studied, but some of them tentatively have been so. A few examples have been given.

## **25.8 Forest Plots for Assessing and Adjusting Baseline Characteristic Imbalance**

Another spin-off of the forest plot methodology is the procedure for assessing and adjusting baseline characteristics in a clinical trial for imbalance. Well et al studied the novel therapeutic compound vandetanib in patients with locally advanced metastatic medullary thyroid carcinoma (J Clin Oncol 2012; 30: 134–141). The data are in the table underneath. Baseline imbalance observed post hoc is not an appropriate reason for including the causing covariate as additional predictor in the primary analysis, and should be interpreted as a random phenomenon. However, with notoriously strong risk factors, a sensitivity analysis including the covariates might be considered. A sensitivity analysis can be defined as a repeat of the primary analysis. A second forest plot is then produced, and with additional covariates novel forest plots for each sensitivity analysis can be performed.



## 25.9 Sensitivity Analysis

With notoriously strong risk factors, a sensitivity analysis including the covariates might be considered. A sensitivity analysis can be defined as a repeat of the primary analysis. A second forest plot is then produced, and with additional covariates novel forest plots for each sensitivity analysis can be drawn. However, if you have strong arguments in favor of a random phenomenon, then skip the sensitivity analysis, like the underneath PONCHO study did with its significant imbalance in genders (*Lancet* 2015; 386: 1261–1268).

Characteristic	Early surgery group	PBD group
Age – yr	64.7 ± 9.5	64.7 ± 10.5
Males – no. (%)	66 (70)	53 (52)
Body-mass index†	24.0 ± 3.1	25.2 ± 3.9
Duration of symptoms – median wks (IQR)	3 (2-6)	3 (2-6)
Weight loss – median kg, (IQR)‡	5 (3-8)	5 (3-10)
Bilirubin level		
Total	151 ± 58.7	154 ± 59.5
Direct	107 ± 49.8	118 ± 57.1
Cause of obstructive jaundice – no. (%)		
Adenocarcinoma	89 (95)	92 (90)
Neuroendocrine tumour	1 (1)	1 (1)
Cystic tumour	1 (1)	3 (3)
Chronic pancreatitis	2 (2)	3 (3)
Adenomatous bile duct polyp	-	1 (1)
Choledocholithiasis	1 (1)	2 (2)

Should we do something with this observation?

## 25.10 Pooled Odds Ratios for Multidimensional Outcome Effects

The outcomes of quality of life (qol) studies is often estimated with multiple item scores, but a better sensitivity of testing may sometimes be obtained with the help of yes/no item responses. As with traditional odds ratio pooling in meta-analyses, yes/no responses of qol may be assessed with odds ratios and pooled odds ratios. An example is given of a study from our group (Application of item response modeling for quality of life assessment, Chap. 7, Clinical Pharmacology Series

16, Zuckschwerdt Verlag New York USA, Zwinderman et al., 1998). 1350 Anginal patients were treated with a novel once daily nitrate compound and the outcome was compared with that of the standard nitrate treatment.

	After 3 months multiple dose therapy	After 3 months once-daily therapy
<b>Mobility difficulties</b>		
walking stairs	429 (33%)	330 (29%)
short distances	354 (27%)	274 (23%)
long distances	587 (46%)	459 (41%)
lifting	464 (36%)	363 (32%)
bending	377 (29%)	338 (29%)
light household work	228 (18%)	194 (17%)
heavy household work	499 (42%)	421 (43%)
profession	95 (15%)	66 (14%)
<b>Pain</b>		
headache	132 (10%)	126 (11%)
backpain	237 (18%)	220 (19%)
pain in upper extremities	258 (20%)	214 (18%)
chest pain	369 (27%)	288 (21%)
<b>Early morning anginal pain</b>		
at wake	81 (6.1%)	52 (4.3%)
when washing	48 (3.6%)	36 (3.0%)
when dressing	55 (4.2%)	41 (3.5%)
<b>Psychological distress</b>		
worrying	282 (21%)	222 (18%)
sleeping problems	332 (25%)	302 (25%)
sombreness	204 (15%)	168 (14%)
irritable	155 (12%)	129 (11%)
Forgetting medication ever	263 (20%)	195 (16%)
Sublingual nitrate use ever	692 (52%)	621 (52%)

The above table gives the numbers of patients with yes answers. The proportions of patients with yes answers equals the risk of a yes answer. From risks odds can be computed according to odds = risk / (1-risk), and odds ratios by comparing the odds of yes in the multiple dose treatment with that of the once daily treatment. Then from all of the odds ratios of the qol domains in the above table. Pooled odds ratios can be calculated very much the same as with meta-analytic pooling of odds ratios (see the Chaps. 1 and 2).

	Odds ratio (95% ci)	p
Mobility difficulties	0.83 (0.76–0.91)	< 0.001
Pain	0.99 (0.84–1.15)	0.85
Chest pain	0.64 (0.48–0.84)	0.001
Early morning anginal pain	0.65 (0.48–0.89)	0.006
Psychological distress	0.87 (0.76–0.99)	0.036
Patient compliance	1.92 (1.37–2.63)	< 0.001
Sublingual nitrate use	0.94 (0.71–1.25)	0.68

Most of the pooled odds ratios were very significant predictors of decreased qol problems with the new treatment. And so the odds ratio procedure is a fine way to assess multidimensional outcome effects of controlled treatment interventions, like quality of life outcomes.

## 25.11 Ratios of Odds Ratios for Subgroup Analyses

The example from the previous section shall be used once more. The odds ratio of mobility difficulties with one daily versus traditional nitrate therapy are given in the previous section. In the above table these odds ratios were calculated for all kinds of subgroups like males, females, young, old patients etc. Then the ratios of these odds ratios of males versus females, young versus old patients etc. were calculated. The table below shows the ratios and their standard errors.

Table I. Patients' characteristics; NYHA = New York Heart Association.

Male gender	848 (63%)
Age (year): mean (SD)	68 (10)
Length (cm): mean (SD)	170 ( 9)
Weight (kg): mean (SD)	76 (13)
Angina class (NYHA)	
I	265 (20%)
II	849 (63%)
III	173 (13%)
IV	14 ( 1%)
Smoking (yes/no)	215 (16%)
Hyperlipidaemia (yes/no)	407 (30%)
Hypertension (yes/no)	400 (30%)
Diabetes mellitus (yes/no)	143 (11%)
Arrhythmia (yes/no)	195 (14%)
Peripheral vessel disease (yes/no)	194 (14%)

The result of this analysis is given below, and was, obviously, very sensitive with almost 30 very significant p-values.

Table IV. Effects of patient characteristics.

	Mobility difficulties	Pain	Early morning anginal pain	Psychological distress	Chest pain	Patient compliance
Gender	1.39 (.13) <sup>c</sup>	1.30 (.17) <sup>c</sup>	1.24 (.43) <sup>b</sup>	1.13 (.17) <sup>c</sup>	0.32 (.27)	-0.10 (.37)
Age	0.014 (.005) <sup>b</sup>	-0.019 (.009) <sup>a</sup>	-0.042 (.021) <sup>a</sup>	-0.045 (.009) <sup>c</sup>	-0.013 (.014)	-0.009 (.017)
NYHA angina class	0.75 (.07) <sup>c</sup>	0.44 (.11) <sup>c</sup>	1.67 (.25) <sup>c</sup>	0.64 (.10) <sup>c</sup>	1.57 (.20) <sup>c</sup>	-0.21 (.24)
Smoking	-0.12 (.17)	0.11 (.21)	1.11 (.64)	0.51 (.21) <sup>a</sup>	-0.27 (.35)	0.76 (.44)
Hyperlipidaemia	-0.03 (.13)	0.14 (.17)	0.11 (.42)	0.26 (.18)	0.25 (.28)	0.04 (.37)
Hypertension	-0.50 (.15) <sup>a</sup>	-0.33 (.17) <sup>a</sup>	-0.18 (.47)	-0.59 (.19) <sup>b</sup>	-0.31 (.29)	0.23 (.38)
Diabetes mellitus	0.34 (.15) <sup>b</sup>	0.05 (.24)	0.96 (.49) <sup>a</sup>	0.31 (.26)	0.25 (.41)	0.05 (.57)
Arrhythmia	0.46 (.15) <sup>b</sup>	0.10 (.23)	0.56 (.44)	0.51 (.23) <sup>a</sup>	1.01 (.36) <sup>b</sup>	0.07 (.50)
Peripheral vessel disease	1.04 (.14) <sup>c</sup>	0.35 (.21)	0.04 (.19)	0.42 (.23)	-0.01 (.35)	0.92 (.49)
Beta blocking medication	-0.10 (.07)	-0.12 (.15)	0.24 (.37)	-0.06 (.15)	0.10 (.25)	0.51 (.33)
CCB medication	0.17 (.11)	0.11 (.15)	0.50 (.32)	0.31 (.15) <sup>a</sup>	0.78 (.25) <sup>b</sup>	0.81 (.33) <sup>b</sup>
Sublingual nitrate use	0.42 (.10) <sup>c</sup>	0.48 (.14) <sup>c</sup>	-0.00 (.14)	0.49 (.14) <sup>c</sup>	0.85 (.23) <sup>c</sup>	-0.53 (.31)
Once daily nitrate therapy	0.04 (.05)	0.11 (.09)	0.05 (.20)	0.05 (.08)	-0.20 (.16)	-0.81 (.20) <sup>c</sup>

CCB = Calcium Channel Blocking; <sup>a</sup>p < 0.05; <sup>b</sup>p < 0.01; <sup>c</sup>p < 0.001.

The same data had already been analyzed using a more traditional approach with outcomes scored on 5 point scales instead of binary, and simple linear regression analysis (Jansen et al., Independent determinants of the beneficial effects of nitrates, Int J Clin Pharmacol Ther 2000; 38: 563–7).

x-variable	regression coefficient (B)	standard error	test (T)	Significance level (P-value)
Age	-0.03	0.04	0.8	0.39
Gender	0.01	0.05	0.5	0.72
Rhythm disturbances	-0.04	0.04	1.0	0.28
Peripheral vascular disease	-0.00	0.01	0.1	0.97
Calcium channel blockers	0.00	0.01	0.1	0.99
beta blockers	0.03	0.04	0.7	0.43
NYHA-classification	-0.08	0.03	2.3	0.02
Smoking	-0.06	0.04	1.6	0.08
body mass index	-0.07	0.03	2.1	0.04
hypercholesterolemia	0.07	0.03	2.2	0.03
hypertension	-0.08	0.03	2.3	0.02
diabetes mellitus	0.06	0.03	2.0	0.05

According to the above table from the linear regression analysis from Jansen et al., only 5 p-values were statistically significant at  $p < 0.05$ , and all of the levels of significance were pretty weak. Obviously, the linear regression model did not at all fit the data as well as did the odds of odds ratio analysis as previously described in meta-analyses. This procedure may, thus, be another pleasant spin-off of meta-analysis methodology.

## 25.12 Conclusion

Modern meta-analyses do more than combine the effect sizes of a series of similar studies. The term “meta” in meta-analysis can be interpreted as “beyond”, and meta-analyses are currently increasingly applied for any analysis beyond the primary analyses of studies. In this chapter also terms like meta-learning, meta-analytic graphing, meta-analytic thinking, meta-cognition, meta-knowledge, meta-strategic knowledge, and awareness of learning processes were addressed. Relevant Spin-offs of meta-analysis methodology were reviewed:

1. Meta-learning
2. Meta-analytic graphing
3. Meta-analytic thinking in writing study protocols and reports
4. Meta-analytic forest plots of baseline patient characteristics
5. Meta-analysis of forest plots of propensity scores
6. Meta-analytic thinking: effect size of important scientific issues other than the main study outcomes
7. Forest plots for adjusting baseline characteristic imbalance
8. Sensitivity analysis
9. Pooled odds ratios for multidimensional outcome effects
10. Ratios of odds ratios for subgroup analyses with item response modeling.

Meta-analysis methodology is used not only for combined outcome effect sizing of series of similar studies. It is also used for other relevant purpose like covering background information of a new study, symmetry assessments of patient characteristics in controlled trials, confounding assessments, sensitivity assessments, item response modeling, and much more.

## Reference

Meta-analysis methodology is reviewed in the Chaps. 1 and 2 of the current edition.

# **Chapter 26**

## **Novel Developments**

### **Pooling Unconventional Outcome Measures**

**Abstract** A condensed overview of modern meta-analytic methodologies as reviewed in the past 25 chapters is given. In addition, meta-analyses pooling unconventional outcome measures are addressed, including meta-analyses of studies tested with analysis of variance, meta-analyses of crossover trials with binary outcomes, bio-equivalence study meta-analyses, and meta-analyses including agenda-driven biases.

#### **26.1 Introduction, Condensed Review of the Past**

Meta-analyses were ‘invented’ in the early 1970s by psychologists, but pooling study results extends back to the early 1900s by statisticians such as Karl Pearson, and Ronald Fisher. In the first years pooling of the data was often impossible due to heterogeneity of the studies. However, after 1995 trials became more homogeneous, due to regulations like standardized protocols and many more requirements (Understanding clinical data analysis, learning statistical principles from published clinical research, Chap. 2, Randomized and observational research, Springer Heidelberg Germany, 2016, from the same authors). In the late 90s several publications concluded that meta-analyses did not accurately predict treatment and adverse effects. The pitfalls were held responsible. Initiatives against them include (1) the Consolidated-Standards-of-Reporting-Trials-Movement (CONSORT), (2) the Unpublished-Paper-Amnesty-Movement of the English journals, and (3) the World Association of Medical Editors’ initiative to standardize the peer review system. Guidelines/checklists for reporting meta-analyses were published like QUOROM (Quality of Reporting of Meta-analyses) and MOOSE (Meta-analysis Of Observational Studies in Epidemiology).

## 26.2 Condensed Review of the Current Edition and Novel Developments

In the current edition many examples of modern methodologies of meta-analysis have been reviewed. The Chap. 5 gave examples of data-analysis using the MetaXL statistical software program of Excel. However, many more software programs for the analysis of meta-data do exist, for example, by SAS, the Cochrane Revman, S-plus, StatsDirect, StatXact, True Epistat, etc. Most of these programs are expensive, but common procedures are also available through Microsoft's Excel and in Excel-add-ins, while many websites offer online statistical analyses for free, including BUGS and R. Leandro's software program (Meta-analysis in medical research. Br Med J books, London UK, 2005). The latter visualizes heterogeneity directly from a computer graph based on Galbraith plots. In the past few years, many new statistical meta-analysis methods have been developed, and all of them have been duly reviewed in the subsequent chapters of this edition.

Chapter 6 showed, that both crossover and parallel-group double blind trials can be analyzed together, although, again, random effect tests should be used to account for the difference in study designs. Also, this chapter demonstrated, that, with large meta-analyses of randomized controlled trials, currently often called the pinnacle of evidence based research, pitfalls are relatively small, for example smaller than 5%, and that they, therefore, need not always be tested.

Chapter 7 showed, that observational studies observational case-control and cohort studies can be simultaneously included in a meta-analysis, sometimes called meta-epidemiological meta-analyses. Random effect tests should assess the difference in study designs. Chap. 7 also showed, that observational plus randomized studies can be included and that the difference in design can be used for sensitivity analysis.

Chapter 9 showed, that in recent years the method of meta-regression brought new insights. For example, group-level instead of patient-level analyses easily fail to detect heterogeneities between individual patients, otherwise, called ecological biases.

Chapter 10 showed, that meta-analysis is also relevant for pooling the performance of diagnostic tests.

Odds ratios are beautiful, but, without an exact confidence interval, they cannot estimate the magnitude of the populations they have been obtained from. Tetrachoric correlation coefficients are helpful for the purpose (Chap. 20).

Studies heterogeneous due to obviously different populations contrast coefficients confidence intervals, generally, better fit meta-analysis models, than Satterthwaite confidence intervals do (Chap. 22).

Chapter 25 addressed the spin-off of meta-analysis methodologies for other forms of clinical data analysis. Forest plots are, for examples, used for the assessment of interaction and confounding in a trial on the outcome.

The current and final chapter will address novel methodologies not yet reviewed, for example, the meta-analysis of ANOVAs (analyses of variance), and of data

mining data sets, equivalence study meta-analyses, agenda-driven meta-analyses including modern phenomena like cherry-picking (selective reporting of successful studies) and ignoring of unsuccessful studies, and the SPIN phenomenon (Chap. 25, section 8). Specific reporting strategies for bolstering your research and/or finances are also spin-offs first recognized by meta-analysts.

## 26.3 Meta-analysis of Studies Tested with Analyses of Variance

In controlled trials usually the differential effect of two treatments is assessed. The t-test is adequate for testing. Meta-analyses of multiple similar studies is performed with pooling. We just take the mean, here the mean difference, of the outcome variable we want to meta-analyze and add up. The data can be statistically tested according to unpaired t-test of the sum of multiple means:

$$t = \frac{\text{mean}_1 + \text{mean}_2 + \text{mean}_3 \dots}{\sqrt{\text{se}_1^2 + \text{se}_2^2 + \text{se}_3^2 + \dots}} \text{ with degrees of freedom} \\ = n_1 + n_2 + n_3 + n_k - k$$

$n_i$  = sample size ith sample,  $k$  = number of samples,  $se$  = standard error of the mean

If the standard deviations are very different in size, e.g., if one is twice the other, then a more adequate calculation of the pooled standard error is as follows. This formula gives greater weight to the pooled  $se$  the greater the samples.

$$\text{Pooled } se = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + \dots}{n_1 + n_2 + \dots - k} \times \left( \frac{1}{n_1} + \frac{1}{n_2} + \dots \right)}$$

Similarly, if the samples are very different in size, then a more adequate calculation of the nominator of  $t$  is as follows.

$$k \left( \frac{\text{mean}_1 n_1 + \text{mean}_2 n_2 + \dots}{n_1 + n_2 + \dots} \right)$$

Alternatively, a weighted mean can be calculated according to

$$\bar{X}_w = \frac{\sum_{i=1}^k w_i x_i}{\sum_{i=1}^k w_i}$$

and its standard error is

$$se(\bar{X}_w) = \left[ \frac{\sum_{i=1}^k (w_i)^2 Var(x_i)}{\left[ \sum_{i=1}^k w_i \right]^2} \right]^{1/2}.$$

where  $x$  = summary statistic of the individual studies (here the individual mean),  $w = 1/se^2$ ,  $se$  = standard error of individual means, var. = variance of individual means. Chi-square tests are used for testing (see also the Chaps. 1 and 2).

If, however, in individual studies, three instead of two treatments are assessed, the pooled t-test will be impossible, and one way analysis of variance or Kruskall-Wallis tests using respectively F-values and chi-square values as test-statistic. For example, in a study of 5 parallel-groups consistent of 8 patients each, the between-group degrees of freedom is  $5 - 1 = 4$  degrees of freedom. The within-group degrees of freedom is  $5 \times 8 - 5 = 35$ .

Let's assume that mean squares (MS-values) and F-value are like shown underneath.

	MS	F	p
within-group	0.15	3.0	$0.05 < p < 0.10$
between-group	0.05		

Then this one way analysis of variance will only produce a trend to a significant difference between the three treatments ( $0.05 < p < 0.10$ ). If a similar study with similar results is available, then you can try and pool the F-values of the two studies.

$$\text{pooled } F = \frac{0.15 + 0.15}{0.05 + 0.05} = 3.0$$

The pooled F-value is identical to that of the individual studies but, now, we will have  $4 + 4 = 8$  degrees of freedom and  $35 + 35 = 70$  degrees of freedom, and, consequently the pooled F-value will produce a p-value  $< 0.05$ . If the assumption of normality is warranted, then a Kruskall-Wallis test will be more adequate (SPSS for starters, Chap. 8, Springer Heidelberg Germany, 2010, from the same authors).

We should add, that any studies with unconventional outcome measures like, e.g., categorical outcomes, can be meta-analyzed similarly to the above approach.

In contrast, regression related effect size estimators like b-values (regression coefficients), phi-coefficients (correlation coefficients for nominal exposure and outcome data) etc. can better be meta-analyzed with the traditional weighted average effect methods (see the Chaps. 1 and 2).

## 26.4 Meta-analyses of Crossover Trials with Binary Outcomes

In Chap. 6 the combined meta-analysis of controlled clinical parallel-group and crossover trials was reviewed, and the weighted average effect approach was used for analysis. The same outcome measure was used in the American meta-analysis of Whelton et al. (Ann Intern Med 2002; 136: 493–503). A slightly different procedure was followed in the English meta-analysis of the Cochrane Library (Ann Intern Med 2008; 148: 30–48) that used ratios rather than differences of average treatment effects in the intervention versus control groups. The procedure are pretty straightforward and The Generic Inverse Variance program of the Cochrane's meta-analysis software program is very helpful for analysis even if limited information from individual studies is available like an estimate and standard error only. With crossovers and paired binary instead of paired continuous data as outcome McNemar's odds ratios (paired odds ratios) instead of the traditional odds ratios must be applied.

With unpaired odds ratios 2 groups, 1 treatment can be analyzed. With McNemar's odds ratios 1 group 2 treatments can be analyzed.

		normotension with drug 1	
		yes	no
normotension with drug 2	yes	(a) 65	(b) 28
	no	(c) 12	(d) 34

Here the  $OR = b/c$ , and the se is not  $\sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}$ , but rather  $\sqrt{\left(\frac{1}{b} + \frac{1}{c}\right)}$ .

$$\begin{aligned} OR &= 28/12 &= 2.33 \\ \ln OR &= \ln 2.33 &= 0.847 \\ se &= \sqrt{\left(\frac{1}{b} + \frac{1}{c}\right)} &= 0.345 \\ \ln OR \pm 2 se & &= 0.847 \pm 0.690 \\ & &= \text{between } 0.157 \text{ and } 1.537, \end{aligned}$$

Turn the ln numbers into real numbers by the anti- $\ln$  button of your pocket calculator.

$$\begin{aligned} &= \text{between } 1.16 \text{ and } 4.65 \\ &= \text{sig diff from } 1.0. \end{aligned}$$

Calculation p-value:  $t = \ln OR/se = 0.847/0.345 = 2.455$ . The bottom row of the t-table produces a p-value of 0.0246, and the two drugs produce, thus, significantly different results at  $p < 0.02$ . For detailed information see Statistics applied to clinical studies 5th edition, Chap. 3, Springer Heidelberg Germany, 2012, from the same authors.

If we wish to meta-analyze two studies with the McNemar's odds ratios of 2.33 and 2.20 and se-values of log OR of 0.345 and 0.400, then the weighted average

effect size can be calculated and tested for level of statistically significant difference from zero. For additional information see Chap. 1.

$$\text{Chi-square} = \frac{\left( \frac{\ln\text{OR}_1}{s_1^2} + \frac{\ln\text{OR}_2}{s_2^2} \right)^2}{\frac{1}{s_1^2} + \frac{1}{s_2^2}}$$

ln = natural logarithm

$$\begin{aligned}\text{Chi-square} &= [\ln 2.33 / 0.345^2 + \ln 2.20 / 0.400^2]^2 / [0(1/0.345^2 + 1/0.400^2)] \\ &= 9.885 \text{ one degree of freedom}\end{aligned}$$

P-value = 0.001667

The meta-analyzed effect size statistics is a lot more precise than that of the single study with a p-value falling from 0.0246 to 0.001667.

## 26.5 Equivalence Study Meta-analysis

Heterogeneity of outcomes is the most troublesome aspect of meta-analyses. And the statistical tests for assessing, whether heterogeneity in a meta-analysis is statistically significant or not, the Cochran's chi-square of Q test, has little power with few studies in the meta-analysis and excessive power with many studies (Chaps. 1 and 2). In addition the random effect heterogeneity test widens confidence intervals, and, if a fixed effect analysis is not significant, then a random effect analysis will almost certainly be insignificant as well (Chap. 4). Fortunately, in the past few years the I-square-statistics has been developed, as a measure for quantity of heterogeneity (Higgins and Thompson, Stat Med 2002; 21: 1539–58). It is complementary to the Q-statistic of the above Q test. The amount of heterogeneity between studies is quantified according to

$$I^2 = 100\% * [Q - (k - 1)] / Q$$

which is interpreted, as the proportion of total variation in study estimates due to heterogeneity, rather than sampling error. It is, like variances and standard deviations, a statistic that is largely independent of sample sizes, which may be a problem to statisticians, but, to non-mathematical clinicians, it should be more appealing, as an estimate of heterogeneity, than a statistically underpowered significance test, like the Cochran Q-test. In addition, it is pretty easy to obtain study number dependent 95% confidence intervals of I-squares and they can be used as a compensation for the lack of an appropriate null-hypothesis test of heterogeneity, as we shall see in the next few lines.

Higgins proposed the H-square value = Q/(k-1), where Q = the Q-statistic which follows a chi-square distribution with k-1 degrees of freedom. It tests the null hypothesis that homogeneity can be rejected.

$$Q = \sum_{i=1}^k w_i (x_i - \bar{X}_w)^2$$

with  $k-1$  degrees of freedom.

H-square measures the amount of heterogeneity.

It can be shown, that

I-square = (H-square - 1) / H-square = proportion variance unexplained.

Thus

$H = \text{excess of } Q \text{ over its degrees of freedom}$

Expected  $Q = k-1$  if no heterogeneity

$H = 1$  means perfect homogeneity.

The frequency distribution of I-square-values from 509 meta-analyses in the Cochrane Library was described by Higgins et al, BMJ 2003; 327: 557–60. Over 50% had I-square = 0%. Only 25% had I-squares over 50%. The underneath table shows the pretty flat frequency distribution of I-square values over 0%. In practice, 50% is often used as a cut-off for heterogeneity.

I-square ( % )	number studies
0	250
0-10	24
10-20	27
20-30	31
30-40	23
40-50	28
50-60	38
60-70	30
70-80	28
80-90	23
90-100	2

This is fine, but, in order to assess, whether this 50% is much or little, we need an interval of equivalence, just like with therapeutic equivalence testing (Statistics applied to clinical studies, Chap. 5, Springer Heidelberg, 2012, from the same authors). We will call the interval of equivalence here the interval of homogeneity, and it is a clinical estimate based on clinically relevant homogeneity, that should be set by the investigators on clinical grounds prior to the completion of the meta-analysis protocol. For example, with randomized-controlled-trial-meta-analyses little heterogeneity is expected (see Chap. 6), unlike with met-analyses of observational studies (see Chap. 8). The intervals may accordingly be set, e.g., respectively

at 95% confidence intervals of the I-square between 0 to 65% and between 0 and 85%.

Examples of randomized controlled trial meta-analyses (set interval 0–65%)

	Measured 95% confidence intervals	homogeneity
meta-analysis 1	40–60%	demonstrated
meta-analysis 2	45–70%	uncertain
meta-analysis 3	70–85%	rejected

Examples of observational study meta-analyses (set interval 0–85%)

	Measured 95% confidence intervals	homogeneity
meta-analysis 1	40–60%	demonstrated
meta-analysis 2	45–70%	demonstrated
meta-analysis 3	70–85%	demonstrated

Higgins and Thompson (Stat Med 2002; 21: 1539–58) provided equations to calculate the 95% confidence intervals of the I-square values (proportion variance unexplained) computed from the data. After log transformation the 95% confidence interval can be calculated from the underneath equations.

$$95\% \text{ confidence} = \text{antilog}(\log H \pm 2 \text{ se}_{\log H})$$

$$\text{se}_{\log H} = 1/2(\log Q - \log(k - 1)) / [(\sqrt{2Q} - \sqrt{2k - 3})]$$

If  $Q < k$ , a slightly different equation must be applied for calculating the se.

$$\text{se}_{\log H} = \left[ 1/(2(k - 2)) \times 1 - 1/\left(3(k - 2)^2\right) \right]^{1/2}.$$

## 26.6 Agenda-Driven Meta-analyses

Many flaws of meta-analyses have been reviewed in the past 25 chapters, except for the most serious one, the agenda-driven meta-analysis. Roseman et al. reported conflict of interest in meta-analyses of pharmacological clinical trials in JAMA 2011; 305: 1008–17. Of 29 meta-analyses only 2 reported funding sources of trials and zero reported industry ties of trial authors. In the BMJ paper of the Beta-blockers in heart failure collaborative group (2016; 353: 1855–60) the beneficial effect of beta-blockers to heart failure patients with a reduced ejection fraction was reported. Like the above 29 meta-analyses, also this one was not adjusted for conflicts of interests. However, the public domain has pretty convincingly provided some of the missing data. Menarini the manufacturer of nebivolol provided an unrestricted grant, Glaxo provided data extraction support, Sankyo (closely monitored by the FDA because of remuneration of physicians and protocol violation and false claims) paid forty million to Medicaid as a settlement. All of the clinical trials

included in the meta-analysis were, of course, industry-funded, and they recruited together over 6000 patients. Conflict of interests remains an important issue in pharmaceutical trials, but there are more “agenda-driven biases”. E.g., Higgins et al. published in BMJ 2011; 343: 5928 a review of the Cochrane “Risk-of-bias tool” written by an international committee of 16 statisticians and epidemiologists (see also Chap. 24, section 6). Matters like

- inadequate allocation concealment,
- blinding of personal and outcome assessors,
- incomplete outcome data,
- selective reporting,

were addressed for the benefit of future investigators.

## 26.7 Hills’ Plurality of Reasoning Statement, Evidence Based Medicine Avant la Lettre

Given the continued presence of risks of biases in clinical research, it is probably wise even in the year 2016 to keep in mind the plurality of reasoning strategies from Professor Bradbury Hills from 1965 (Proceedings Roy Soc Med 1965; 58: 295–300).

1. Strength of associations.
2. Consistency of results.
3. Specificity of variables.
4. Temporality of effects.
5. Dose-response patterns.
6. Biological plausibility.
7. No conflict with other relevant knowledge.
8. Controlled experiments.
9. Analogy with other accepted cause effect relationships.

The above nine rules remain a superior strategy for addressing evidence based medicine even in the current era of big data and wonderful software for the purpose of analyses (Machine learning in medicine a complete overview, Springer Heidelberg Germany, 2016, from the same authors).

## 26.8 Conclusion

In this chapter a condensed review of meta-analyses in the past and of modern meta-analyses as reviewed in this edition are given. Meta-analyses of studies with unconventional outcomes like analysis of variance f-values, paired binary odds

ratios, I square values of previous meta-analyses, conflict of interest as meta-analysis outcome. Finally the famous Hills plurality of reasoning statement was humbly recognized.

## Reference

To readers requesting more background, theoretical and mathematical information of the computations given in this edition, several textbooks complementary to the current production and written by the same authors are available: Statistics applied to clinical studies 5th edition, 2012, Machine learning in medicine a complete overview, 2015, SPSS for starters and 2nd levelers 2nd edition, 2015, Clinical data analysis on a pocket calculator 2nd edition, 2016, Understanding clinical data analysis from published research, 2016, all of them edited by Springer Heidelberg Germany.

# Index

## A

- Adjusted Heterogeneity without Overdispersion, 63
- Agenda-driven bias, 40
- Agenda-Driven Meta-Analyses, 306–307
- Alternative Methods for diagnostic meta-analyses, 133
- Antihypertensive effect of potassium, a meta-analysis, vii
- Approximation Methods for Computing Correlation Coefficients from Odds Ratios, 238–239
- Assessing Separate Effects of Predictors on One Outcome Adjusted for the Other, 217
- Automatic Data Mining in SPSS Modeler, 269
- Automatic data mining programs, v
- Awareness of learning processes and thinking skills, v

## B

- Bayesian networks, 145
- Benefits and Criticisms of Meta-analyses, 19–21
- Beta-blockers and heart failure, a meta-analysis, vi
- Big data, v
- Big research data, v
- Binary Outcome Data, Fixed Effect Analysis, 55–56
- Binary Outcome Data, Meta XL Free Meta-Analysis Software, 67–75
- Binary Outcome Data, Odds Ratio, 33

## Binary Outcome Data, Random Effect Analysis, 56–58

- Binary Outcome Data, Relative Risk, 32–33
- Binary Outcome Data, Risk Difference, 32
- Binary Outcome Data, Survival Data, 33–34
- Bivariate approach to meta-analysis, vii
- Bottom-up Bayesian network, 151
- Bucher network, 155
- BUGS, 39

## C

- Case-control studies, 39
- Challenging the Exchangeability Assumption, 243, 244
- Chi-square statistics, 56
- Christmas tree plot, 15
- Clinical data analysis on a pocket calculator 2nd edition, v
- Cochran Q test, 36
- Cochrane Collaborators, 44
- Cochrane Library, vii
- Cochrane Revman, 24
- Cohort studies, 8, 39
- Collaterals and deaths and re-infarctions, vii
- Combined Meta-Analysis of Different Classes of Study Designs, 93
- Comprehensive Meta-analysis, 24
- Condensed Review of the Current Edition, 300–301
- Condensed Review of the Past, 299
- Confidence Intervals Methods for Indirect Comparisons, 245–247
- Confirmatory clinical trials, vi
- Confounding, vi

- C**
- Consolidated-Standards-of-Reporting-Trials-Movement (CONSORT), 40
- Constructing Summary Receiver Operated Curves 114, 132–133
- Contingency table, 6
- Continuous Outcome Data, Coefficient of determination  $R^2$  or  $r^2$  and Its Standard Error, 31–32
- Continuous Outcome Data, Correlation Coefficient ( $R$  or  $r$ ) and Its Standard Error, 29–31
- Continuous Outcome Data, Fixed Effect Analysis, 58–59
- Continuous Outcome Data, Mean and Standard Deviation, 25
- Continuous Outcome Data, Online Meta-Analysis Calculator, 64–66
- Continuous Outcome Data, Random Effect Analysis, 60–61
- Continuous Outcome Data, Regression Coefficient and Standard Error, 27–28
- Continuous Outcome Data, Strictly Standardized Mean Difference (SSMD), 26–27
- Continuous Outcome Data, Student's T-Value, 28–29
- Contrast coefficients, vi
- Contrast coefficient model, fixed effect meta-analysis, 252
- Contrast coefficient model, random effect meta-analysis, 253–254
- Contrast Coefficients Meta-Analysis, 249–259
- Contrast Testing Using One Way Analysis of Variance on SPSS Statistical Software, 257–258
- Convenience Samples and Other Limitations, 79
- Correlation Coefficients as a Replacement of Odds Ratios, 233
- Correlation coefficients to z transformations, vi
- Cumulative logit link function, 161
- D**
- Dersimonian and Laird, 17
- Diagnostic Odds Ratios (DORs), 129
- Dose Response Meta-Analyses, 167
- Drug research, vi
- E**
- Effect sizes of a series of similar studies, v
- Ensemble Learning in SPSS Modeler Example, 196
- Ensemble Learning with SPSS Modeler, Example, 207–215
- Ensembled Accuracies, 205–216
- Ensembled Correlation Coefficients, 195–204
- Epidemiologists, vi
- Equivalence Study Meta-Analysis, 304–306
- Event Analysis in Patients with Collateral Coronary Arteries, 103
- Evolutionary operations (EVOP) meta-analysis of three studies, 262
- EVOP, First Study, 263
- EVOP, Second Study, 263–265
- EVOP, Third Study, 265–266
- Example of contrast coefficients model, 250–252
- Example of diagnostic odds ratios, 129–131
- Example 1, the Potassium Meta-analysis, 44–45
- Example 2, the Calcium Channel Blocker Meta-Analysis, 45
- Example 3, the Large Randomized Trials Meta-Analyses, 45–46
- Example 4, the Diabetes and Heart Failure Meta-Analysis, 46
- Example 5, the Adverse Drug Effect Admissions and the Type of Research Group Meta-Analysis, 46–47
- Example 6, the Coronary Events and Collaterals Meta-analysis, 47
- Example 7, The Diagnostic Meta-Analysis of Metastatic Lymph Node Imaging, 47
- Example 8, The Homocysteine and Cardiac Risk Meta-Analysis, 48
- Exploring the Effects of Small Changes in Experimental settings, 261
- extras.springer.com, vi
- F**
- First pitfalls, 15
- Fixed-effect estimate, 24
- Fixed effect model, 17
- Forest Plots for Assessing and Adjusting Baseline Characteristic Imbalance, 292–293
- Frequentists' Methods for Indirect Comparisons, 244–245
- Frequentists' Networks, 145
- Funnel plot, 70, 74
- G**
- Galbraith plots, 39
- General Framework, 23
- General Loglinear Modeling, 180–183

**H**

- Handling categories, vii  
Heterogeneity, 2, 16  
Heterogeneity Rather than Homogeneity of Studies as Null-Hypothesis, 270–272  
Higher order of thinking, v  
Hills' causality rules, 20  
Hill's Evidence Based Medicine Avant la Lettre, 307  
Hills plurality of reasoning statement, 308  
Homocysteinemia and coronary artery disease, meta-analysis, vi  
How to Perform a Meta-Analysis, 5–8  
Hypothesis against control observations, 48  
Hypothesis-driven meta-analyses, 40

**I**

- Innovative meta-analyses, v  
Interaction, vi  
Invert variance heterogeneity (IVhet) method, 64  
 $I^2$  square, 84

**J**

- JAMA, vi

**K**

- Konstanz information miner (KNIME), vi

**L**

- Leandro's software, 39  
Learning process, v  
Less Noise but More Risk of Overfit, 195, 205  
lnOR (log odds ratio), 55  
Lumley networks, 155

**M**

- Machine learning in medicine a complete overview, v  
Mantel Haenszel summary chisquare test, 14  
Mathematical Framework, 23–41  
Mean variance, 26  
META XL from Excel, vi  
Meta XL Free Meta-Analysis Software, 67–75  
Meta XL Free Meta-Analysis Software for Excel, 67–75  
Meta-Analyses of Randomized Controlled Trials (RCTs), 79–91  
Meta-analyses of diagnostic tests, vi  
Meta-Analyses, Novel Developments, 298  
Meta-Analyses with Direct and Indirect Comparisons, 243–248  
Meta-Analyses with Multivariate Assessments, 217–231  
Meta-Analysis and Random Effect Analysis, 51–62  
Meta-Analysis in a Nutshell, 1–22  
Meta-analysis of calcium channel blockers in chronic heart failure, vii  
Meta-Analysis of Diagnostic Studies, 127–134  
Meta-Analysis of Forest Plots of Propensity Scores, 287–290  
Meta-analysis of heart failure with diabetes mellitus, vii  
Meta-Analysis of Observational Plus Randomized Studies, 93–100  
Meta-analysis of Observational Studies, 101–114  
Meta-analysis of Observational Studies in Epidemiology, 40  
Meta-Analysis of Studies Tested with Analyses of Variance, 301–302  
Meta-analysis of Three EVOP Studies, 266–267  
Meta-Analysis Software Programs, 63  
Meta-analysis with diagnostic tests, vi  
Meta-Analysis with Evolutionary Operations (EVOPs), 261–267  
Meta-Analysis with General Loglinear Models, 177–183  
Meta-Analysis with Heterogeneity as Null-Hypothesis, 269–277  
Meta-analysis with loglinear models, 177–183  
Meta-analysis with R, vi  
Meta-Analysis with Variance Components, 185–193  
Meta-Analysis with Weighted Least Square Regressions, 177  
Meta-analytic forest plots, v  
Meta-Analytic Forest Plots of Baseline Patient Characteristics, 284–287  
Meta-analytic Graphing, 280–282  
Meta-analytic spin-off, effect size of “outcome reporting bias”, 291  
Meta-analytic spin-off, effect size of blinding, 291  
Meta-analytic spin-off, effect size of placebo effects on the studies’ outcome, 291  
Meta-analytic spin-off, effect size of prospective nature of prospective, 291  
Meta-analytic spin-off, effect size of randomization, 291  
Meta-analytic spin-off, effect sizes of optimistic abstracts, and conclusion based on subgroups, 291

Meta-Analytic Thinking and Other Spin-Offs of Meta-Analysis, 279–298  
 Meta-Analytic Thinking in Writing Study Protocols and Reports, 282–283  
 Meta-Analytic Thinking: Effect Size Assessments of Important Scientific, 290–292  
 Meta-Analyzing Rare Diseases, 101  
 Meta-cognition, v  
 Meta-knowledge, v  
 Meta-learning, 279  
 Meta-Meta-Analysis, 135–143  
 Meta-Meta-Analysis for Meta-Learning Purposes, 141–142  
 Meta-Meta-Analysis for Re-assessment of the Pitfalls, 136–141  
 Meta-regression, 39, 115–126  
 Meta-regression confounding, 117–119  
 Meta-regression exploratory purpose, 116–117  
 Meta-regression, binary Outcome, 121–124  
 Meta-regression, confounding, 122  
 Meta-regression, Continuous Outcome, 115–120  
 Meta-regression, exploratory Purpose, 121–122  
 Meta-regression, interaction, 119–120, 123–124  
 Modern meta-analyses, v  
 Multistage regression, vii  
 Multinomial probability distribution, 161  
 Multiple Categorical Outcome and Predictor Variables, 157  
 Multiple Regression as an Alternative to Subgroup Analyses, 115  
 Multivariate meta-analysis, example 1, 218–221  
 Multivariate meta-analysis, example 2, 222–226  
 Multivariate meta-analysis, example 3, 226–231

**N**  
 Network Meta-Analysis, 145–155  
 New Developments, 39–40  
 Newton’s rules of the scientific method, 44  
 Nonmathematical professionals, v  
 Null Hypothesis Testing of Linear Contrasts, 255–256

**O**  
 Observational plus randomized studies: improved information from combined assessments, 99–100  
 Observational plus randomized studies:pooled Results, 96–98  
 Observational plus randomized studies: summary statistics, 95–96  
 Observational plus randomized studies heterogeneity Assessments, 98  
 Observational plus randomized studies publication Bias Assessments, 98–99  
 Observational plus randomized studies robustness Assessments, 99  
 Odds, 6  
 Odds Ratios, 6, 234, 238–239  
 Old and New Style Random Effect Analysis, 51  
 Online meta-analysis calculator from www.healthstrategy.com, 63  
 Online Meta-Analysis Calculators, 63–64  
 Original Meta-Analyses, 136–141

**P**  
 Pearson Goodness of fit test, 168  
 Peto’s z-test, 14  
 Pinnacle of Science or an Error-Ridden Exercise, 1  
 Pitfalls, Heterogeneity, 35–37  
 Pitfalls, Lack of Sensitivity, 38  
 Pitfalls, Publication Bias, 34–35  
 Pocket calculator methods, vi  
 Pocket Calculator One-Way Analysis of Variance (ANOVA) of Linear Contrasts, 256–257  
 Poisson modeling, 180  
 Pooled Odds Ratios for Multidimensional Outcome Effects, 294–296  
 Pooling Unconventional Outcome Measures, 299  
 Post hoc analyses, 21  
 Primary analysis of studies, v  
 Primer in statistics, vii  
 Principles of Linear Contrast Testing, 254–255  
 Prior hypothesis, 8  
 Probit Regression, 167  
 Prospective open evaluation studies, 102  
 Prospective open evaluation studies, Event Analysis of Iatrogenic Hospital Admissions, 108

- Prospective open evaluation studies,  
    Heterogeneity, and Lack of Robustness,  
        110–112
- Prospective open evaluation studies,  
    Meta-Regression, 112–113
- Prospective open evaluation studies,  
    Meta-regression Analysis, 106–107
- Prospective open evaluation studies, Overall  
    Results, 110–112
- Prospective open evaluation studies, Pooling  
    Results, 102
- Prospective open evaluation studies,  
    Publication bias, 104, 109
- Prospective open evaluation studies, tests for  
    Heterogeneity and Robustness,  
        104–106
- Prospective open evaluation studies, the  
    scientific method, 103, 108–109
- Psychologists have invented meta-analyses in  
    1970, vi
- Publication bias, 2
- Q**
- Q statistic, 36
- Quality of reporting of meta-analyses  
    (QUOROM), 40
- Quasi Likelihood Modeling, 63
- R**
- R software, vi
- Random effect model, 17
- Random-effects estimate, 24
- Random intercepts meta-analysis, 165
- Ratios of Odds Ratios for Subgroup Analyses,  
    296–297
- RCTs:Confirming the Scientific Question, 85
- RCTs:Handling Multiple Outcomes, 87–88
- RCTs:Large Meta-Analyses Without Need for  
    Pitfall Assessment, 88–90
- RCTs:Multiple Outcomes, 85–87
- RCTs:Single Outcomes, 81–85
- Real Data Example of Indirect Comparisons,  
    247
- Reformulate your scientific question into  
    a hypothesis, 43
- Regression Coefficients and Correlation  
    Coefficients as Replacement of Odds  
        Ratios, 234–237
- Risk, 6
- Reducing Unexplained Variance, Increasing  
    Accuracy, 185
- Reformulate your scientific question into  
    a hypothesis, 48
- Relative risk (RR), 6
- Robustness, 2
- S**
- SAS, 24
- Satterthwaite confidence intervals, 39
- Scientific method, 43
- Scientific Rigor and Scientific Method, Two  
    Different Things, 43
- Scientific Rigor, Rule 1, 8–10
- Scientific Rigor, Rule 2, 10–11
- Scientific Rigor, Rule 3, 11
- Scientific Rigor, Rule 4, 12–15
- Second Pitfall, 16–19
- Sensitivity analysis, 19, 292
- Shift from a Single to Multiple Meta-Analyses,  
    135
- Shift from P-Values to Effect Sizes of Any  
    Scientific Issue, 279
- Software, 257–258
- Sound Clinical Arguments and Scientific  
    Question, 94–95
- Special Method for Assessing Random Effect  
    Heterogeneity, 249
- SPIN phenomenon, 40
- S-plus, 39
- Springer Heidelberg Germany, v
- SPSS, 24
- SPSS data files, vi
- SPSS for starters and 2nd levelers 2nd edition,  
    v
- SPSS Modeler, vi
- SPSS Modeler, Accuracy Assessment  
    of Decision Tree Output, 276
- SPSS Modeler, Beneficial Effects of the  
    Treatments, Histograms, 272–273
- SPSS Modeler, Beneficial Effects of  
    Treatments, 275–276
- SPSS Modeler, Beneficial Effects  
    of Treatments, Clusters, 273
- SPSS Modeler, Beneficial Effects of  
    Treatments, Decision trees, 275–276
- SPSS Modeler, Frequency Distribution of the  
    Treatments, 272
- SPSS Modeler, Network of Causal  
    Associations Displayed as a Web, 274
- SPSS' work bench for automatic data mining,  
    vi
- Stata, 24
- StatsDirect, 39

- Step by step analyses, vi  
Strict inclusion criteria, 9  
Summary Receiver Operating (ROC) Curves, 127
- T**  
Tau square, 74  
Tetrachoric correlation, 39  
Tetrachoric Correlation Coefficients from an Odds Ratio, 239–241  
Third Pitfall, 19  
Thompson's I square, 52  
Thorough search of trials, 9  
Top-down Bayesian network, 150  
Transforming Odds Ratios into Correlation Coefficients, 233–242  
True Epistat, 39
- U**  
Understanding clinical data analysis from published research, v  
Uniform guidelines for data analysis, 9
- Unpublished-Paper-Amnesty-Movement, 40  
Updated methodologies, vi
- V**  
Variance, 26  
Variance components, example 1, 185–187  
Variance components, example 2, 187–189  
Variance components, example 3, 189–192  
Visualizing Heterogeneity, 53–54
- W**  
Weighted average effect, 24  
Weighted average effect analysis, 14  
Weighted Multiple Linear Regression, 177–180  
Working Papers with Emphasis on Entire Data Coverage, 23  
World Association of Medical Editors, 40