Abteilung Biometrie und Medizinische Dokumentation, Universität Ulm
Leiter: Prof. Dr. Wilhelm Gaus
Betreuer: PD Dr. Josef Högel

CENTER WEIGHTING FACTORS AND TREATMENT EFFECTS IN MULTICENTER CLINICAL TRIALS

Application of the Analysis of Variance in Medicine



Dissertation zur Erlangung des Doktorgrades der Humanbiologie
(Dr. biol. hum.)
der Medizinischen Fakultät der Universität Ulm

Michaela Carla Glasbrenner
aus Heilbronn

Ulm, December, 2002

Amtierender Dekan: Prof. Dr. Reinhard Marre

1. Berichterstatter: PD Dr. Josef Högel

2. Berichterstatter: Prof. Dr. Hans Wolff

Tag der Promotion: 14. Februar 2003

Contents

1 Introduction **1**

2 Fixed center effects **7**

2.1 Introduction 7

2.2 Model formulation 9

2.3 Estimation and test of overall treatment effect 21

2.3.1 Analysis I 27

2.3.2 Analysis II - Cochran, Mantel-Haenszel Estimator 30

2.3.3 Analysis III - Yates Approach 35

2.3.4 No interactions 37

2.3.5 Equal number of patients receiving the two treatments at each center 38

2.4 Comparison of Analyses I to III 40

3 Random center effects **45**

3.1 ANOVA Model 46

3.2 General Mixed Model 48

3.3 Estimation of overall treatment effect 52

3.3.1 DerSimonian and Laird Method 57

3.3.2 Likelihood Model Approach 60

3.4 Test of overall treatment effect 63

3.5 Meta-Analysis 64

4 Simulation of treatment effects **66**

4.1 Comparison of fixed and random center effects 66

4.2 Simulation of data from fixed and mixed models 70

4.2.1 Results 72

4.3 Weight engendering functions 77

4.3.1 Fixed center effect 77

4.3.2 Random center effect 79

4.3.3 Results of sensitivity analyses 81

5 The standard center **91**

5.1 Divergence measures 91

5.1.1 Kullback-Leibler Number 91

5.1.2 Matusita Distance 92

5.2 On wisdom before the trial 94

5.2.1 The new weighting factor 95

5.2.2 Possibility one 96

5.2.3 Possibility two 97

5.3 Additional concomitant variables 98

5.4 Examples 100

5.5 Future work 108

6 Summary and Conclusions **109**

A Tables concerning the Transpulmin Study **115**

List of Abbreviations and Symbols

ANOVA	Analysis of variance
BLUE	Best linear unbiased estimator
BLUP	Best linear unbiased predictor
χ^2_n	χ^2 distribution with n degrees of freedom
CMH	Cochran, Mantel-Haenszel (estimator)
$Cov(.,.)$	Covariance between two random variables
df	Degrees of freedom
$E(.)$	Expected value of a random variable
$E(x \mid y)$	Conditional expectation of x for given y
$f_X(x)$	Probability density function of the random variable X
$f_{Y\|X}(x,y)$	Conditional density function of Y for given X
GLM	Generalized linear model
ICH	International conference on harmonization
MINQUE	Minimum norm quadratic unbiased estimation
MIVQUE	Minimum variance quadratic unbiased estimation
ML	Maximum likelihood
MLE	Maximum likelihood estimator
MSE	Mean squared error
$N(E, Var)$	Normal distribution of a random variable with expected value E and variance Var
REML	Restricted maximum likelihood
SAS	Statistical analysis system
SRP	Study represented population
SS	Sums of squares
SSE	Error sums of squares
SSTO	Total sums of squares
SSTR	Treatment sums of squares
$t_n(.)$	Student's t distribution with n degrees of freedom

$t_{n,p}$	100*p%-quantile of the t distribution with n degrees of freedom
UMVUE	Uniformly minimum variance unbiased estimator
$Var(.)$	Variance of a random variable
X^-	Generalized inverse of X

List of Tables

2.1 Fixed weighting factors and derived terms 27

2.2 The different hypotheses (SAS handbook [39]) 27

2.3 Fixed weighting factors with unequal and equal sample sizes 38

4.1 Distribution of patients of first simulation study 70

4.2 Significance level of first simulation study 72

4.3 Power of first simulation study 74

4.4 Distribution of patients of second simulation study 84

5.1 Sample sizes per center and treatment of Transpulmin Study 101

5.2 Center deviations of Transpulmin Study 102

5.3 Weighting factors of Transpulmin Study 102

5.4 Characteristics of Transpulmin Study 103

5.5 Significance level of new approach based on data from a mixed effects model 105

5.6 Power of new approach based on data from a mixed effects model . . 105

A.1 Age distribution of the Transpulmin Study 116

A.2 Smoking distribution of the Transpulmin Study 117

A.3 Distribution of degree of cold at the beginning of the Transpulmin Study 118

List of Figures

4.1 Decision on fixed or random approach 67

4.2 Significance level functions of first simulation study 81

4.3 Power functions of first simulation study 83

4.4 Significance level functions of second simulation study 85

4.5 Power functions of second simulation study 86

4.6 Significance level functions of Transpulmin Study 87

4.7 Power functions of Transpulmin Study 88

5.1 Decision on type of model 106

Chapter 1

Introduction

Multicenter trials - as the name implies - are trials at any stage of the clinical development which are

> *"conducted at more than one distinct center, where the data collected from theses centers are intended to be analyzed as a whole" (Hsin-Tsung et al. 1998 [27]).*

A study is usually performed at multiple sites because a single center is not capable of accruing the necessary sample size in a reasonable time interval. This may be the case when the disease under investigation is rare, the outcome of the therapy uncommon or the anticipated treatment effect too small (Horwitz et al. 1996 [26]).

Therefore, a first advantage of multicenter studies compared to single center studies is the relatively large number of patients that can be recruited within a short period of time. Secondly, the results obtained can be generalized to a wider spectrum of patients and centers because of the "possibility of recruiting subjects from a wider population and of administering the medication in a broader range of clinical settings" (Guideline E9 of the International Conference on Harmonization on "Statistical Principles for Clinical Trials"). Generalization will be increased even further

if a trial is conducted in a number of different countries. It will then be called a multinational multicenter trial (ICH E9 guideline). However, differences in culture, medical practice, regulatory requirements and drug management emerge and need to be eliminated (Hsin-Tsung et al. 1998 [27]). As a third advantage, multicenter studies allow the participation of small centers and enforce the cooperation and exchange of knowledge among the investigators in the different centers.

Along with the many advantages of multicenter trials there are also some disadvantages which have to be mentioned. If the responses vary substantially among the centers, it will not be advisable to pool the observations into a single analysis. Providing the centers are not too small the data of each hospital should be analyzed separately. An overall conclusion might then be missing or might not be clear. The analysis would seem unfinished. Berry (1990 [2]) stated in this respect: "If statistics cannot make a single inference from multivarious sources, there is something wrong with statistics". This key problem of multicenter trials will be addressed at length in this writing.

A further difficulty of multicenter trials is the statistical analysis in cases of varying sample sizes at the different centers as well as varying sample sizes between the treatment groups. In general, multicenter studies are evaluated with analysis of variance (ANOVA) models. When the sample sizes are unequal, the *regular* analysis of variance formulas are no longer appropriate (Neter et al. 1996 [36]). More complex formulas based on the general linear test approach need to be applied. Furthermore, unequal center sizes will increase the problem of deducing an overall conclusion out of responses which vary substantially among the centers. Appropriate weights for the responses of the centers need to be found. This is the reason why the ICH E9 guideline suggests that a "good design should generally aim to achieve the same distribution of subjects to treatments within each center and trials should also avoid excessive variation in the number of subjects per center". In practice, the different centers are usually not able to recruit the same number of patients since there are larger and smaller centers. Different centers may also have a different patient

clientele. In general, four different types of multicenter designs can be distinguished (Kaufmann 2001 [30]):

1. a small number of centers with many patients at each
2. a moderate number of centers with a moderate number of patients at each
3. a large number of centers with a small number of patients at each
4. a moderate to large number of centers with widely varying numbers of patients at each.

The sample sizes of the treatment groups are often unequal at a center but due to the randomization process they do not vary as extremely as the sample sizes between the centers. Furthermore, potential bias may be introduced where a few centers have very slow enrollment. The sponsor or trial manager may wish to take the inefficient centers out of the study and increase the required sample size of the more active sites (Hsin-Tsung 1998 [27]).

A further problem of multicenter trials is associated with their evaluation based on analysis of variance models. Different types of ANOVA models exist and a consensus on the most appropriate one has not been reached yet. In order to describe the different models, a specific multicenter trial is considered which aims to explore the efficacy of a new pharmaceutical in comparison to placebo. It is further anticipated that the efficacy of the pharmaceutical is measured by a quantitative score. Concerning data evaluation after the end of the study, the simplest model will investigate the influence of the verum and the placebo treatment on the score. The response, i.e. the score, will therefore be written as a function of the qualitative factor "treatment". Since only one factor is included in the model the analysis is termed Analysis I.

With regard to multicenter trials it is also interesting to know if the participating centers influence the score results. One center may in general yield higher scores

than a second center - for the verum as well as for the placebo group. In this case a center effect is present. Moreover, if the centers achieve different scores and if the differences between the verum and the placebo scores vary among the centers, an interaction between treatment and center will exist. Hence it seems reasonable to expand the model. The score will then not only depend on the treatment but also on the centers (Analysis II) and sometimes additionally on the treatment-by-center interaction (Analysis III). There is disagreement whether a "treatment-by-center interaction" effect should be part of the model or not.

Next to the discussion about the number of factors included in the model, there is further disagreement on the type of the factors. A factor can be either fixed or random. When interest is only in the effect of the specific factor levels chosen, e.g. the "verum" and the "placebo" level of the factor "treatment" or the different centers participating in the trial, the factor will be fixed. In contrast, it will be random if the factor levels are a sample from a larger population of potential levels and if inferences are desired about the whole population (Neter et al. 1996 [36]). For example, if the factor "center" is chosen as random, interest will not only be in the participating clinics, but in the population of clinics out of which the participating clinics are assumed to be a random sample.

There is agreement that the treatment effect should be a fixed factor. However, the discussion about the type of the center effect and hence of the treatment-by-center interaction is still ongoing. Models which consist of fixed factors only are called fixed effects models. The corresponding analyses are Analyses I to III. In contrast, models with at least one fixed and one random factor are denoted as mixed effects models. They are also called as the "random approach" since the factor "center" is random and since there is only one type of analysis available for mixed effects models. The discussion about fixed and random center effects therefore leads to a comparison of fixed and mixed effects models.

With regard to fixed effects models, the varying number of factors included in the analysis is responsible for the treatment effects of the centers being weighted

differently. There exist three different weights (weights I to III) according to models with a treatment effect only (Analysis I), models with a treatment and a center effect (Analysis II), and models incorporating a treatment, a center and a treatment-by-center interaction effect (Analysis III). Whether to include or not a treatment-by-center interaction with a treatment and a center effect is therefore tantamount to a choice between weights II and III. For mixed effects models, only one weighting factor exists. It is termed as the "random weight".

Multicenter trials and associated problems are described in a phletora of articles and textbooks. Literature can commonly be divided into two groups. Group one focuses on the generalized linear model theory, where multicenter trials are special cases thereof. On the other hand, publications can be found describing only parts of the complex topic in an applied manner, i.e. important for the concrete evaluation of medical trials. Next to an abundance in terminology and notation which causes confusion, connections between the two groups are difficult to find. Furthermore, concerning the various approaches to data evaluation, no consensus has yet been reached. Thus, the first aim of this writing is the comprehensive presentation of all aspects of the topic as well as the connections between them. Current approaches will be investigated at length and their advantages and disadvantages will be compared. Many new investigations including the introduction of a completely new weighting factor will be performed in order to bring forward the design and analysis of multicenter trials. At the end a recommendation on the most suitable approach to use will be given.

The outline of the writing is as follows. In Chapter 2, fixed effects models will be presented both from a theoretical and from a practical side. Unequal sample sizes and quantitative normally distributed responses are assumed. A terminology is suggested which describes ANOVA models in general but which can afterwards be applied with ease to the specific types of analyses available. The structure of Chapter 3 is similar to the one of Chapter 2. Chapter 3 is devoted to models with a random center effect. Along with the test statistic of the traditional mixed effects

model for the investigation of an overall treatment effect, the recently introduced test statistic by Hartung et al. (2001 [21] [22]) will be described and examined. The statistic is also based on a mixed effects model and will be called as the "Hartung statistic".

The peculiarities of models with a fixed and a random center effect will be summarized at the beginning of Chapter 4. Simulation studies will be performed in order to compare the different approaches directly via various terms related to hypothesis testing. After the evaluation of current methods used in practice as well as recent ones found in the literature, sensitivity analyses based on further simulation studies will be conducted. For a first analysis, the two commonly employed weights of the fixed effects model (weights II and III) are considered to be extreme choices which can be connected via certain functions. Further weights can consequently be derived from such functions which offer a compromise between the extremes. Next to an exponential function, a linear function will be introduced briefly. The dependence of estimates and test decisions on the "intermediate" weights will be explored and graphically depicted in order to assess the robustness of estimates and decisions of the usual weights.

A second sensitivity analysis based on models with a random center effect will be performed. Mixed effects models do not only consider the variability of responses within a center but also the variability between the centers. Since both variances are usually not known they have to be estimated. The sensitivity analysis examines the influence of an imprecise moment estimate of the between-center-variability on the outcome.

Finally in Chapter 5, a new weighting factor will be introduced. It will rest upon a fixed effects model and will incorporate the distance of a center from a predefined standard center with the help of a divergence measure called the Matusita distance. The standard center will be specified by major patient characteristics. By this means, patient heterogeneity between the centers can be quantified and included into a fixed effects model.

Chapter 2

Fixed center effects

2.1 Introduction

In this writing the problem under investigation is the comparison of two independent groups of patients, each receiving a certain treatment. Patients of a group A are assumed to obtain the active treatment while patients of a group B are assumed to be either treated with placebo or with an active control. For simplicity purposes, it will be anticipated that patients of group B obtain placebo.

In order to demonstrate a treatment effect in a randomized controlled clinical trial, a certain sample size per treatment group is required. In practice, it is not always possible for one clinic to recruit enough patients in a reasonable time frame. As a standard solution, several clinics will take part in the study.

In the entire writing, a quantitative and normally distributed response is assumed as efficacy measure of the verum and the placebo treatment. With regard to a single-center trial, the treatment mean of the verum group can be easily compared with the treatment mean of the placebo group. The difference of both is the estimated treatment effect. However, in case of several hospitals, data may be affected by the treatments as well as by the hospitals. This is due to center-specific variations

in many factors. Goldberg and Koury (1990 [15]) write that "when a rating scale is used, for example, the investigator's overall level of efficacy scores may differ by center, trial procedures and equipment may be implemented differently, and equipment may be calibrated not in the exact same manner". In addition to different methods of measurement, study personnel as well as patients may vary between the centers. All of these factors are reasons why there are often substantial discrepancies in the outcomes of the participating hospitals. On this account the factor "center" is usually considered as a stratification variable so that the treatment effects are estimated at each center. An overall treatment effect estimate is then obtained in averaging over the estimated center effects. The overall (across-center) estimate is thus adjusted for center (Goldberg and Koury 1990 [15]). In case the patients are randomly assigned to the treatment groups within a center, the center-specific effects are however equally distributed over the treatment groups within the centers (Salsburg 1990 [38]).

As previously mentioned there are three commonly used approaches for models with fixed effects. The first method only distinguishes between patients receiving treatment A and patients receiving treatment B. The fact of different participating centers is not considered. The second approach weights the centers with regard to the variation of their data, whereas the third approach considers all centers as equally important. It assigns the same weight to all centers. In some special cases the three methods coincide. An investigation of multicenter clinical trials published in PubMed revealed, that a high proportion of multicenter trials is still evaluated with the first approach and is therefore performed without considering any center information. The three approaches and their particularity will be described in detail in the following sections.

2.2 Model formulation

It is assumed that the influence of two main factors ("center", "treatment") as well as their interaction on a response variable Y is to be examined. The first factor "center" will be investigated at C different levels and the second factor "treatment" at 2 levels which represent the two possible treatments. The number of patients receiving at center c (c=1, ... , C) treatment t (t=A, B) will be denoted as n_{ct}.

Let Y be a response variable which is quantitative. It is further assumed that all observations are independent. Y_{ctk} be the response of the kth patient (k=$1, 2, \ldots, n_{ct}$) receiving treatment t (t=A or B) at center c (c=1, ... , C). Suppose additionally that the responses Y_{ctk} are normally distributed with mean μ_{ct} and variance σ^2. Therefore, $Y_{ctk} \sim N(\mu_{ct}, \sigma^2)$.

For the investigation of treatment, center and/or treatment-by-center interaction effects, a two-factor analysis of variance model (ANOVA model) is appropriate. There are many different versions of this model depending on the number of factors included, depending on whether the factors are chosen as random or as fixed, and whether or not an equal number of observations is available for each factor level combination (Graybill 1976 [17]). The factor "treatment" is chosen as a fixed effect because interest is only in the two particular factor levels "verum" and "placebo". The reason for a fixed center effect is that "in practice, the selection of centers is not random although it is not generally subject to any known biases related to treatment effects"(Jones [29]). A random center effect will be discussed in Chapter 3.

In the entire writing the special case of empty cells will not be considered. Each center is therefore assumed to recruit at least one patient for each treatment group. The models are also called 2-way crossed classification models with all-cells-filled data.

For the formulation of the model for the underlying problem, a vast amount of varying terminology exists. A short presentation of the most commonly used formulations and notations will be given (e.g. Neter et al. 1996 [36], Searle 1971 [42], 1987 [44]).

- The appropriate **cell means model** of a two-factor analysis of variance with interaction (and fixed effects only) is denoted as

$$Y_{ctk} = \mu_{ct} + e_{ctk}\,, \tag{2.1}$$

with

  - the cell means μ_{ct} being parameters that represent the average effect of the cth level of the factor "center" and the tth level of the factor "treatment"
  - the random error terms e_{ctk} being independent $\mathrm{N}(0, \sigma^2)$ distributed
  - c=1, ... , C; t=A, B; k=1, ... , n_{ct} .

- The **factor effects model** represents a second possibility. It is obtained by splitting up the parameter μ_{ct} of model (2.1) in a general mean $\mu^{\star}_{..}$, a center main effect $\alpha^{\star}_{c}$, a treatment main effect $\beta^{\star}_{t}$ and possible treatment-by-center interactions $(\alpha\beta)^{\star}_{ct}$

$$\mu_{ct} = \mu^{\star}_{..} + \alpha^{\star}_{c} + \beta^{\star}_{t} + (\alpha\beta)^{\star}_{ct}\,.$$

The following model results

$$Y_{ctk} = \mu^{\star}_{..} + \alpha^{\star}_{c} + \beta^{\star}_{t} + (\alpha\beta)^{\star}_{ct} + e_{ctk}\,, \tag{2.2}$$

where

  - the general mean $\mu^{\star}_{..}$ is a constant,
  - $\alpha^{\star}_{c}$ are constants representing the effect of the cth level of the factor "center"

- $\beta_t^\star$ are constants representing the effect of the tth level of the factor "treatment"
- $(\alpha\beta)_{ct}^\star$ are constants representing the interaction effect when "center" is at the cth level and "treatment" is at the tth level.

The factor effects model (2.2) has more parameters than observed cell means. This is called overparameterization. In order to obtain unique parameter estimates at a later point in time, additional restrictions are required. They are the reason why subsequent model (2.3) is also often called a $\sum$-restricted model. The parameters of the restricted model are presented without an asterisk to be able to distinguish them from the parameters of the unrestricted model (2.2). The restricted model is given by

$$Y_{ctk} = \mu_{..} + \alpha_c + \beta_t + (\alpha\beta)_{ct} + e_{ctk}\,, \tag{2.3}$$

where

- α_c are constants subject to the side condition $\sum_{c=1}^{C} \alpha_c = 0$
- β_t are constants subject to the side condition $\sum_{t=A}^{B} \beta_t = 0$
- $(\alpha\beta)_{ct}$ are constants subject to the restrictions

  $\sum_c (\alpha\beta)_{ct} = 0 \quad$ t=A, B

  $\sum_t (\alpha\beta)_{ct} = 0 \quad$ c=1,..., C
- e_{ctk} are independent N(0, σ^2)
- c=1, . . . , C; t=A, B; k=1, . . . , n_{ct} .

Restricted and unrestricted models can be merged into one another. Out of the parameters of an unrestricted model, new parameters can be developed (e.g. Scheffé 1964 [41], Searle 1971 [42], Searle 1987 [44])

$$
\begin{aligned}
\mu_{..} &= \mu^{\star}_{..} + \frac{\sum_c \alpha^{\star}_c}{C} + \frac{\sum_t \beta^{\star}_t}{2} + \frac{\sum_c \sum_t (\alpha\beta)^{\star}_{ct}}{2C} \\
\alpha_c &= \alpha^{\star}_c - \frac{\sum_c \alpha^{\star}_c}{C} \\
\beta_t &= \beta^{\star}_t - \frac{\sum_t \beta^{\star}_t}{2} \\
(\alpha\beta)_{ct} &= (\alpha\beta)^{\star}_{ct} - \frac{\sum_c (\alpha\beta)^{\star}_{ct}}{C} - \frac{\sum_t (\alpha\beta)^{\star}_{ct}}{2} + \frac{\sum_c \sum_t (\alpha\beta)^{\star}_{ct}}{2C} \, ,
\end{aligned}
$$

so that the side conditions and restrictions of model (2.3) hold in any case. For example,

$$
\sum_c \alpha_c = \sum_c \alpha^{\star}_c - \sum_{c=1}^{C} \frac{\sum_c \alpha^{\star}_c}{C} = 0 \, .
$$

The full rank model (2.3) is obtained.

Interest is sometimes in center or treatment effects which are weighted in certain ways. In this case it is more convenient to utilize the restrictions

$$
\begin{aligned}
\sum_c w_c \, \alpha_c &= 0 \\
\sum_t v_t \, \beta_t &= 0 \\
\sum_c w_c \, (\alpha\beta)_{ct} &= 0 \quad t = A, \, B \\
\sum_t v_t \, (\alpha\beta)_{ct} &= 0 \quad c = 1, \ldots, C \, ,
\end{aligned} \tag{2.4}
$$

where w_c are center weights and v_t are treatment weights with $\sum_{c=1}^{C} w_c = 1$ and $\sum_{t=A}^{B} v_t = 1$.

- As a third possibility, the problem can also be expressed in terms of a **regression approach** (e.g. Neter et al. 1996 [36], page 893)

$$\begin{aligned} Y_{ctk} &= \mu_{..} + \alpha_1 X_{c,t,k,1} + \ldots + \alpha_{C-1} X_{c,t,k,C-1} + \beta_A X_{c,t,k,C} \\ &\quad + (\alpha\beta)_{1,1} X_{c,t,k,1} X_{c,t,k,C} + \ldots + (\alpha\beta)_{C-1,1} X_{c,t,k,C-1} X_{c,t,k,C} + e_{ctk} \ , \end{aligned}$$

with

$$X_{c,t,k,1} = \begin{cases} 1 & \text{if center 1} \\ 0 & \text{otherwise} \\ -1 & \text{if center C} \end{cases}$$

.

.

.

$$X_{c,t,k,C-1} = \begin{cases} 1 & \text{if center C-1} \\ 0 & \text{otherwise} \\ -1 & \text{if center C} \end{cases}$$

$$X_{c,t,k,C} = \begin{cases} 1 & \text{if treatment A} \\ -1 & \text{if treatment B} \end{cases} .$$

The remaining parameters are defined as

- $\alpha_c = \mu_{c.} - \mu_{..}$ for c=1, ... , C-1
- $\beta_A = \mu_{.A} - \mu_{..}$ and
- $(\alpha\beta)_{cA} = \mu_{cA} - \mu_{c.} - \mu_{.A} + \mu_{..}$.

The dot notation represents unweighted averages of the cell means μ_{ct} (e.g. $\mu_{c.} = \frac{1}{C} \sum_{c=1}^{C} \mu_{ct}$). Any other parameters are not required. The model has full rank. Therefore, no constraints are necessary.

- Last but not least, the above problem can be described in a very general way with the matrix notation of the **generalized linear model (GLM) theory**

$$\underset{n_{\cdot\cdot}\times 1}{\underline{Y}} = \underset{n_{\cdot\cdot}\times 2C}{\underline{\underline{X}}}\ \underset{2C\times 1}{\underline{\vartheta}} + \underset{n_{\cdot\cdot}\times 1}{\underline{\epsilon}}, \tag{2.5}$$

where $n_{\cdot\cdot} = \sum_{c=1}^{C}\sum_{t=A}^{B} n_{ct}$, $\underline{\vartheta}$ represents the parameter vector and $\underline{\underline{X}}$ the design matrix. More precisely, the vector and matrices of e.g. the cell means model are

$$\begin{pmatrix} \underline{Y}_{1An_{1A}} \\ \underline{Y}_{1Bn_{1B}} \\ \underline{Y}_{2An_{2A}} \\ \underline{Y}_{2Bn_{2B}} \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_{CAn_{CA}} \\ \underline{Y}_{CBn_{CB}} \end{pmatrix} = \begin{pmatrix} \underline{1}_{n_{1A}} & & & & & & & & \underline{0} \\ & \underline{1}_{n_{1B}} & & & & & & & \\ & & \underline{1}_{n_{2A}} & & & & & & \\ & & & \underline{1}_{n_{2B}} & & & & & \\ & & & & \cdot & & & & \\ & & & & & \cdot & & & \\ & & & & & & \cdot & & \\ & & & & & & & \underline{1}_{n_{CA}} & \\ \underline{0} & & & & & & & & \underline{1}_{n_{CB}} \end{pmatrix} \begin{pmatrix} \mu_{1A} \\ \mu_{1B} \\ \mu_{2A} \\ \mu_{2B} \\ \cdot \\ \cdot \\ \cdot \\ \mu_{CA} \\ \mu_{CB} \end{pmatrix} + \begin{pmatrix} \underline{e}_{1An_{1A}} \\ \underline{e}_{1Bn_{1B}} \\ \underline{e}_{2An_{2A}} \\ \underline{e}_{2Bn_{2B}} \\ \cdot \\ \cdot \\ \cdot \\ \underline{e}_{CAn_{CA}} \\ \underline{e}_{CBn_{CB}} \end{pmatrix}$$

with

$$\underline{Y}_{ctn_{ct}} = \begin{pmatrix} Y_{ct1} \\ Y_{ct2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{ctn_{ct}} \end{pmatrix} \qquad \underline{1}_{n_{ct}} = \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}_{n_{ct}\times 1} \qquad \underline{e}_{ctn_{ct}} = \begin{pmatrix} e_{ct1} \\ e_{ct2} \\ \cdot \\ \cdot \\ \cdot \\ e_{ctn_{ct}} \end{pmatrix},$$

for c=1, ... , C and t=A, B.

For the estimation of the factor effects of any of these models there are two commonly used approaches: the **least squares method** and the **maximum likelihood method** (Dobson 1990 [8]). The first is based on minimizing the sum of the squared deviations of the responses from their expected values, while the second maximizes (the logarithm of) the likelihood function. The likelihood function is algebraically the same as the joint density of a random sample. The difference is that for the joint density function the Y_{ctk} are variables and the parameters are fixed. With regard to the likelihood function, the parameters are variable and the observations are fixed. For many generalized linear models, the maximum likelihood estimates can only be obtained numerically by means of an iterative procedure. This is due to non-linear equations which are often obtained after differentiating the log-likelihood function. Two commonly used iterative procedures are the Newton-Raphson method and the method of scoring.

For the problem under investigation, the least squares method and the maximum likelihood method lead to the same requirement for the estimation of the parameters

$$\begin{aligned}(\underline{Y} - \underline{\underline{X}}\,\underline{\vartheta})^T (\underline{Y} - \underline{\underline{X}}\,\underline{\vartheta}) &= \sum_{c=1}^{C}\sum_{t=A}^{B}\sum_{k=1}^{n_{ct}}(Y_{ctk} - \mu_{ct})^2 \\ &= \sum_{c=1}^{C}\sum_{t=A}^{B}\sum_{k=1}^{n_{ct}}(Y_{ctk} - \mu_{..} - \alpha_c - \beta_t - (\alpha\beta)_{ct})^2 \longrightarrow min\,.\end{aligned}$$

The factor effect estimates $\hat{\underline{\vartheta}}$ are then obtained as solutions of the **normal equations** $\underline{\underline{X}}^T\underline{\underline{X}}\,\hat{\underline{\vartheta}} = \underline{\underline{X}}^T\underline{y}$. In case the matrix has full rank, the estimates are unique and are given by

$$\hat{\underline{\vartheta}} = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\underline{\underline{X}}^T\underline{y} \qquad \text{with} \qquad \hat{\underline{\vartheta}} \sim N(\underline{\vartheta}, \sigma^2(\underline{\underline{X}}^T\underline{\underline{X}})^{-1})\,. \tag{2.6}$$

Concerning the properties of the parameter estimates, it is a well known fact that the

parameters are consistent[1], sufficient[2], unbiased[3] and do have minimum variance[4]. The **best linear unbiased estimator** (BLUE)[5] of the mean μ_{ct} is

$$\hat{\mu}_{ct} = \hat{\mu}_{..} + \hat{\alpha}_c + \hat{\beta}_t + \widehat{(\alpha\beta)}_{ct} = \overline{Y}_{ct.} = \frac{1}{n_{ct}} \sum_{k=1}^{n_{ct}} Y_{ctk} \; . \tag{2.7}$$

Since the μ_{ct} are estimable so is any linear function of the estimable μ_{ct}'s (Searle 1971 [42]). In contrast, the effects β_t or any linear combination of them, e.g. $\beta_A - \beta_B$, are not individually estimable because of the interaction terms $(\alpha\beta)_{ct}$ and the α-main effects in μ_{ct}. The same holds for the α_c effects, $\mu_{..}$ or $(\alpha\beta)_{ct}$ respectively (Searle 1971 [42]).

Therefore, the cell means model (μ_{ct} - model) is much more simple than the factor effects model. There is no confusion as to which functions are estimable and also which restrictions are to be used in order to obtain unique parameter estimates (Searle 1971 [42], 1987 [44]).

In the special case of a restricted model with **balanced** data ($n_{ct} \equiv n \; \forall c \; \forall t$), the normal equations simplify and the elements (μ_{ct}, $\mu_{..}$, α_c, β_t, $\alpha\beta_{ct}$) become all estimable (Searle 1971 [42]). The (BLUE) parameter estimates of a balanced model are obtained from the following general equations

$$\begin{aligned}
\alpha_c &= \mu_{c.} - \mu_{..} && c = 1, \ldots, C \\
\beta_t &= \mu_{.t} - \mu_{..} && t = 1, \ldots, T \\
\alpha\beta_{ct} &= \mu_{ct} - \mu_{c.} - \mu_{.t} + \mu_{..} && c = 1, \ldots, C \quad t = 1, \ldots, T \; ,
\end{aligned}$$

[1] "An estimator $\hat{\vartheta}$ is a consistent estimator of ϑ if: $\lim_{n\to\infty} P(|\hat{\vartheta} - \vartheta| \geq \epsilon) = 0$ for any $\epsilon > 0$".

[2] "An estimator $\hat{\vartheta}$ is a sufficient estimator of ϑ if the conditional joint probability function of the sample observations, given $\hat{\vartheta}$, does not depend on the parameter ϑ".

[3] "An estimator $\hat{\vartheta}$ is an unbiased estimator of ϑ if: $E[\hat{\vartheta}] = \vartheta$".

[4] "An estimator $\hat{\vartheta}$ is a minimum variance estimator of ϑ if for any other estimator $\hat{\vartheta}^\star$: $Var[\hat{\vartheta}] \leq Var[\hat{\vartheta}^\star] \forall \; \hat{\vartheta}^\star$". (Neter et al. 1996 [36])

[5] "The best linear unbiased estimator (BLUE) of a parameter is that linear function of the observations which is unbiased for the parameter and which, among all such linear unbiased functions, has smallest variance". (Searle 1987 [44])

by replacing μ_{ct} and its linear combinations

$$\begin{aligned}
\mu_{c.} &= \frac{1}{T}\sum_{t=1}^{T}\mu_{ct} \\
\mu_{.t} &= \frac{1}{C}\sum_{c=1}^{C}\mu_{ct} \\
\mu_{..} &= \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\mu_{ct}
\end{aligned}$$

with the corresponding linear functions of $\hat{\mu}_{ct} = \overline{Y}_{ct.}$. It results

$$\begin{aligned}
\hat{\mu}_{..} &= \overline{Y}_{...} \\
\hat{\alpha}_{c} &= \overline{Y}_{c..} - \overline{Y}_{...} \\
\hat{\beta}_{t} &= \overline{Y}_{.t.} - \overline{Y}_{...} \\
\widehat{\alpha\beta}_{ct} &= \overline{Y}_{ct.} - \overline{Y}_{c..} - \overline{Y}_{.t.} + \overline{Y}_{...} \quad .
\end{aligned}$$

Since balanced data analysis is much easier to perform and to interpret than unbalanced data analysis, there is the wish to deal unbalanced data in the same way. This can sometimes be realized in case the "unbalanced data is not too far removed from being balanced" (Searle 1971 [42], p. 361). Usually minor modifications are applied to the data in order to be able to perform a balanced analysis. These will be, first of all, the estimation of missing observations if many observations are the same. Secondly, if the number of observations in the different cells do not vary too much from each other, it may seem reasonable to set some data aside so that the cells are equally filled. Since the definition of what is "not too far removed" remains unsolved, a further discussion of this topic has not been taken into consideration.

The evaluation of unbalanced data is more complicated than the evaluation of balanced data because the least squares approach does no longer yield easy solutions (Neter et al. 1996 [36]). Furthermore, the main ANOVA property that the treatment

sums of squares (SSTR)[6] and the error sums of squares (SSE)[7] add to the total sums of squares (SSTO)[8] does not hold in the unbalanced case. Hence, SSTO $\neq$ SSTR + SSE. The generalized linear test approach is then usually required. The regression model formulation is often employed with it. However, other model formulations are as well possible, as long as a reduced model is fitted next to the full model. The reduced model is obtained from the full model $\underline{Y} = \underline{\underline{X}}\,\underline{\vartheta} + \underline{\epsilon}$ in applying certain restrictions. It is necessary because of the lack of orthogonality of the ANOVA decomposition (Neter et al. 1996 [36]).

If interest is in the investigation of a treatment effect of model (2.1) and if the participating centers are regarded as being of equal importance, the corresponding nullhypothesis to be tested will be

$$H_0: \quad \mu_{.A} = \mu_{.B} \tag{2.8}$$

with $\mu_{.t} = \frac{1}{C}\sum_{c=1}^{C} \mu_{ct}$ for t=A, B. In matrix notation the nullhypothesis of equal treatment means can be paraphrased as

[6]Notation of balanced case:

$SSTR = n \sum_{c=1}^{C} \sum_{t=A}^{B} (\overline{Y}_{ct.} - \overline{Y}_{...})^2$, where $\overline{Y}_{ct.} - \overline{Y}_{...}$ is the deviation of the estimated treatment mean around the overall mean

[7]$SSE = \sum_{c=1}^{C} \sum_{t=A}^{B} \sum_{k=1}^{n} (Y_{ctk} - \overline{Y}_{ct.})^2$, where $Y_{ctk} - \overline{Y}_{ct.}$ is the deviation around the estimated treatment mean

[8]$SSTO = \sum_{c=1}^{C} \sum_{t=A}^{B} \sum_{k=1}^{n} (Y_{ctk} - \overline{Y}_{...})^2$, where $Y_{ctk} - \overline{Y}_{...}$ is the total deviation

$$H_0 \; : \; \underset{s\times p}{\underline{\underline{D}}} \; \underset{p\times 1}{\underline{\vartheta}} = \; \underset{s\times 1}{\underline{h}} \tag{2.9}$$

$$H_0 \; : \quad \underset{1\times 2C}{\left[\frac{1}{C}, -\frac{1}{C}, \frac{1}{C}, -\frac{1}{C}, \ldots, \frac{1}{C}, -\frac{1}{C}\right]} \underset{2C\times 1}{\begin{pmatrix} \mu_{1A} \\ \mu_{1B} \\ \mu_{2A} \\ \mu_{2B} \\ . \\ . \\ . \\ \mu_{CA} \\ \mu_{CB} \end{pmatrix}} = \underset{1\times 1}{0} \; . \tag{2.10}$$

The matrix D is of size $1 \times 2C$ so that s=1 and p=2C.

If the treatment means of the different centers are not of equal importance, other weights w_c with $\sum_{c=1}^{C} w_c = 1$ might be used leading to

$$H_0 : \quad \mu_{.A} = \mu_{.B} \tag{2.11}$$

with $\mu_{.t} = \sum_{c=1}^{C} w_c \, \mu_{ct}$ for t=A, B.

The test statistic for investigating the hypothesis of equal treatment means (2.9) consists of different error sums of squares (Neter et al. 1996 [36]). The error sums of squares of the full (F) model $\underline{Y} = \underline{\underline{X}}\,\underline{\vartheta} + \underline{\epsilon}$ (2.5) are obtained as the squared deviations of the responses Y_{ctk} from their own estimated means $\hat{Y}_{ctk}$

$$\begin{aligned} SSE(F) &= \sum_{c=1}^{C}\sum_{t=A}^{B}\sum_{k=1}^{n_{ct}} (Y_{ctk} - \hat{Y}_{ctk})^2 = (\underline{Y} - \underline{\underline{X}}\,\underline{\hat{\vartheta}}_F)^2 \\ &= (\underline{Y} - \underline{\underline{X}}\,\underline{\hat{\vartheta}}_F)^T (\underline{Y} - \underline{\underline{X}}\,\underline{\hat{\vartheta}}_F) = \underline{Y}^T\underline{Y} - \underline{\hat{\vartheta}}_F^T\,\underline{\underline{X}}^T\,\underline{Y}\,, \end{aligned}$$

where $\hat{\underline{\vartheta}}_F$ is the vector of parameter estimates of the full model calculated according to (2.6). In contrast, the reduced model (R) is defined as

$$\underline{Y} = \underline{\underline{X}}\,\underline{\vartheta} + \underline{\epsilon} \quad \text{with condition} \quad \underline{\underline{D}}\,\underline{\vartheta} = \underline{h} \; .$$

The error sums of squares for this model are given by

$$SSE(R) = (\underline{Y} - \underline{\underline{X}}\,\hat{\underline{\vartheta}}_R)^T (\underline{Y} - \underline{\underline{X}}\,\hat{\underline{\vartheta}}_R) = \underline{Y}^T\underline{Y} - \hat{\underline{\vartheta}}_R^T\,\underline{\underline{X}}^T\,\underline{Y}$$

with

$$\hat{\underline{\vartheta}}_R = \hat{\underline{\vartheta}}_F - (\underline{\underline{X}}^T\,\underline{\underline{X}})^{-1}\underline{\underline{D}}^T\left(\underline{\underline{D}}(\underline{\underline{X}}^T\,\underline{\underline{X}})^{-1}\underline{\underline{D}}^T\right)^{-1}\left(\underline{\underline{D}}\,\hat{\underline{\vartheta}}_F - \underline{h}\right) \; .$$

As shown in many textbooks, the general linear test statistic for investigating the hypothesis $\underline{\underline{D}}\,\underline{\vartheta} = \underline{h}$ then is

$$\begin{aligned} F &= \frac{\frac{SSE(R)-SSE(F)}{df_{red}-df_{full}}}{\frac{SSE(F)}{df_{full}}} = \frac{\frac{SSE(R)-SSE(F)}{(n_{..}-p+s)-(n_{..}-p)}}{\frac{SSE(F)}{n_{..}-p}} = \frac{\frac{SSE(R)-SSE(F)}{s}}{MSE(F)} \\ &= \frac{\frac{\left(\underline{\underline{D}}\,\hat{\underline{\vartheta}}_F-\underline{h}\right)^T\left(\underline{\underline{D}}(\underline{\underline{X}}^T\,\underline{\underline{X}})^{-1}\underline{\underline{D}}^T\right)^{-1}\left(\underline{\underline{D}}\,\hat{\underline{\vartheta}}_F-\underline{h}\right)}{s}}{MSE(F)} \; . \end{aligned} \tag{2.12}$$

A difference of s degrees of freedom results when the degrees of freedom of the reduced model df_{red} are subtracted from the degrees of freedom of the full model df_{full}. For hypotheses (2.8) and (2.11), s has the value 1. The upper denominator of formula (2.12) therefore equals 1. The error sums of squares of the full model divided by their degrees of freedom are denoted as the error mean squares (MSE) of the full model.

Instead of a two-factor analysis of variance model, a randomized complete block design could be applied with the factor "center" as blocking factor. In a randomized block design, the aim is the combination of the experimental units (in this case the

patients) in blocks. The experimental units within a block should be as homogeneous as possible, whereas the different blocks should be as heterogeneous as possible with respect to the response variable (Neter 1996 [36]). This proceeding reduces the error variability. However, the model formulation as well as the mathematical calculations underlying the block design will equal a two-way analysis of variance model if the blocking factor ("center") will be put as additional factor into the model (next to the factor "treatment").

2.3 Estimation and test of overall treatment effect

The terminology - especially the vector and matrix notation - presented so far is very useful with regard to the error sums of squares and the formulation of test statistics. However, for a deeper investigation of treatment and center effects, subsequent notation is preferred. Model (2.3) will from now on be used as the underlying model if not stated otherwise.

The treatment effect for center c is the difference of the mean responses of the two treatments

$$\theta_c = \mu_{cA} - \mu_{cB} \,.$$

Hence, the estimator of the treatment effect for center c is calculated as

$$\begin{aligned} \hat{\theta}_c &= \hat{\mu}_{cA} - \hat{\mu}_{cB} \overset{(2.7)}{=} \frac{1}{n_{cA}} \sum_{k=1}^{n_{cA}} Y_{cAk} - \frac{1}{n_{cB}} \sum_{k=1}^{n_{cB}} Y_{cBk} \\ &= \bar{Y}_{cA.} - \bar{Y}_{cB.} \,, \end{aligned} \tag{2.13}$$

since the best linear unbiased estimator of the mean μ_{ct} has been shown to be $\frac{1}{n_{ct}} \sum_{k=1}^{n_{ct}} Y_{ctk}$, t=A, B; c=1,...,C.

Let σ^2 be the common variance of each of the C centers. The variance of the treatment effect estimator for center c then is

$$Var(\hat{\theta}_c) = \sigma^2 \left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}} \right), \tag{2.14}$$

which will be denoted as v_c^2. Assuming normality, the estimator of the treatment effect for center c has a normal distribution with mean θ_c and variance v_c^2

$$\hat{\theta}_c \quad \sim \quad N(\theta_c, \sigma^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}})) = N(\theta_c, v_c^2) \; .$$

In order to obtain an overall treatment effect, the treatment effects of the centers need to be combined. This will be performed by weighting the center treatment means and by summing up the center differences

$$\theta \;=\; \sum_{c=1}^{C} (\, w_{cA}\, \mu_{cA} - w_{cB}\, \mu_{cB}\,) \,. \tag{2.15}$$

The w_{ct} represent weighting factors subject to the constraint $\sum_{c=1}^{C} w_{ct} = 1$ for t=A, B. They will be further specified in later sections.

Let $\hat{\theta}$ be the estimator of the overall difference in treatment means

$$\begin{aligned} \hat{\theta} \;&=\; \sum_{c=1}^{C} (\, w_{cA}\, \hat{\mu}_{cA} - w_{cB}\, \hat{\mu}_{cB}\,) \\ &=\; \sum_{c=1}^{C} \left(w_{cA}\, \bar{Y}_{cA.} - w_{cB}\, \bar{Y}_{cB.} \right) , \end{aligned} \tag{2.16}$$

with the following expected value and variance

$$\begin{aligned} E(\hat{\theta}) \quad &= \quad \sum_{c=1}^{C} E\left(w_{cA}\, \bar{Y}_{cA.} - w_{cB}\, \bar{Y}_{cB.} \right) \\ &= \quad \sum_{c=1}^{C} w_{cA}\, \mu_{cA} - w_{cB}\, \mu_{cB} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{c=1}^{C} w_{cA}(\mu_{..} - \alpha_c - \beta_A - (\alpha\beta)_{cA}) - w_{cB}(\mu_{..} - \alpha_c - \beta_B - (\alpha\beta)_{cB}) \\
&= \beta_A - \beta_B + \sum_{c=1}^{C} (\, \alpha_c(w_{cA} - w_{cB}) + w_{cA}\,(\alpha\beta)_{cA} - w_{cB}\,(\alpha\beta)_{cB}\,)
\end{aligned}
\tag{2.17}
$$

$$
\begin{aligned}
Var(\hat{\theta}) &= Var\left(\sum_{c=1}^{C}(\, w_{cA}\,\bar{Y}_{cA.} - w_{cB}\,\bar{Y}_{cB.}\,)\right) \\
&\overset{\overline{Y}_{ct.}\ ind.\forall c}{=} \sum_{c=1}^{C} w_{cA}^2\, Var\,(\,\overline{Y}_{cA.}\,) + w_{cB}^2\, Var\,(\,\overline{Y}_{cB.}\,) \\
&= \sum_{c=1}^{C}(\, w_{cA}^2\,\frac{\sigma^2}{n_{cA}} + w_{cB}^2\,\frac{\sigma^2}{n_{cB}}\,) \\
&= \sigma^2 \sum_{c=1}^{C}\left(\frac{w_{cA}^2}{n_{cA}} + \frac{w_{cB}^2}{n_{cB}}\right) \\
&= \sigma^2\left(\frac{1}{W_A} + \frac{1}{W_B}\right) = \sigma^2\left(\frac{W_A + W_B}{W_A\,W_B}\right) .
\end{aligned}
\tag{2.18}
$$

For convenience purposes, W_t is defined as $W_t = \left(\sum_{c=1}^{C} \frac{w_{ct}^2}{n_{ct}}\right)^{-1}$ for t=A, B according to Scheffé (1964 [41]).

The test statistic of investigating the hypothesis for an overall treatment effect is given by

$$
Z = \frac{\hat{\theta}}{\sqrt{Var(\hat{\theta})}} = \frac{\sum_{c=1}^{C}\,(\, w_{cA}\,\bar{Y}_{cA.} - w_{cB}\,\bar{Y}_{cB.}\,)}{\sqrt{\sigma^2 \sum_{c=1}^{C}\left(\frac{w_{cA}^2}{n_{cA}} + \frac{w_{cB}^2}{n_{cB}}\right)}} .
\tag{2.19}
$$

Under the nullhypothesis of equal treatment means, Z is $N(0,1)$ distributed while it is $N(\frac{\theta}{\sqrt{Var(\theta)}}, 1)$ distributed under the alternative.

In case the variance σ^2 is not known beforehand, the mean squared errors of formula

(2.12) are employed as unbiased estimator of the residual error variance

$$MSE = \hat{\sigma}^2 = \frac{SSE}{df} = \frac{\sum_{c=1}^{C}\sum_{t=A}^{B}\sum_{k=1}^{n_{ct}}(Y_{ctk} - \overline{Y}_{ct.})^2}{n_{..} - 2C} \qquad (2.20)$$

$$= \frac{\sum_{c=1}^{C}\sum_{t=A}^{B}(n_{ct}-1)\,S_{ct}^2}{n_{..} - 2C} \;,$$

where

$$n_{..} = \sum_{c=1}^{C}\sum_{t=A}^{B} n_{ct}$$

and

$$S_{ct}^2 = \frac{1}{n_{ct}-1}\sum_{k=1}^{n_{ct}}(Y_{ctk} - \overline{Y}_{ct.})^2 \quad .$$

The estimator is unbiased because of the well known fact that

$$E(MSE) = \sigma^2$$

under model (2.3). In the denominator of (2.19), the variance will now be replaced by its estimate. A t-statistic with $n_{..} - 2\,C$ degrees of freedom results. In squaring the t-statistic and applying subsequent notations (Scheffé 1964 [41]) for c=1, . . . , C and t=A, B

$$\tilde{Y}_{.t.} = \sum_{c=1}^{C} w_{ct}\,\overline{Y}_{ct.} = \sum_{c=1}^{C}\frac{w_{ct}}{n_{ct}}\sum_{k=1}^{n_{ct}} Y_{ctk}$$

$$\tilde{Y}_{...} = \frac{\sum_{t=A}^{B} W_t\,\tilde{Y}_{.t.}}{\sum_{t=A}^{B} W_t} \qquad \text{where} \qquad W_t = \left(\sum_{c=1}^{C}\frac{w_{ct}^2}{n_{ct}}\right)^{-1} \quad , \qquad (2.21)$$

the following F statistic will be obtained

$$F = \left(\frac{\tilde{Y}_{.A.} - \tilde{Y}_{.B.}}{\sqrt{\hat{\sigma}^2\,(\frac{1}{W_A} + \frac{1}{W_B})}}\right)^2 = \frac{\frac{W_A\,W_B}{W_A+W_B}\,(\tilde{Y}_{.A.} - \tilde{Y}_{.B.})^2}{\hat{\sigma}^2}$$

$$\underset{=}{\text{Searle [44]}} \quad \frac{\sum_{t=A}^{B} W_t\,(\tilde{Y}_{.t.} - \tilde{Y}_{...})^2}{\hat{\sigma}^2} \ . \tag{2.22}$$

As stated at great length in Scheffé (1964 [41]) but also in the writings of Elston et al. (1964 [9]), Searle (1971 [42]) and Neter et al. (1996 [36]), the sum $\sum_{t=A}^{B} W_t(\tilde{Y}_{.t.} - \tilde{Y}_{...})$ is the appropriate weighted sum of squares for testing the hypothesis

$$H_0: \quad \underline{\underline{D}}\,\underline{\vartheta} = \underline{h}$$

$$H_0: \quad [w_{1A}, -w_{1B}, w_{2A}, -w_{2B}, \ldots, w_{CA}, -w_{CB}] \begin{pmatrix} \mu_{1A} \\ \mu_{1B} \\ \mu_{2A} \\ \mu_{2B} \\ . \\ . \\ . \\ \mu_{CA} \\ \mu_{CB} \end{pmatrix} = 0 \tag{2.23}$$

of equal treatment means. Yates (1934 [54]) was the first who introduced the weighted sums of squares analysis with weighting terms W_t being in inverse proportion to the variance of the $\tilde{Y}_{.t.}$ terms. Formula (2.22) can be rewritten in a more general form as

$$\begin{aligned} F &= \frac{\sum_{t=A}^{B} W_t\,(\tilde{Y}_{.t.} - \tilde{Y}_{...})^2}{\hat{\sigma}^2} = \frac{\text{Weighted SS of treatment effect}}{MSE} \\ &= \frac{\frac{SSTreatment_w}{1}}{MSE} = \frac{\frac{SSTreatment_w}{df}}{MSE} \ . \end{aligned} \tag{2.24}$$

The connection of the F statistic of formulas (2.22) and (2.24) with the general linear test statistic (2.12) is now quite apparent and will be further specified at a later point in time.

In practice, the investigation of the above mentioned hypothesis is commonly performed with the "Statistical Analysis System" (SAS Institute, Cary, NC, U.S.A), especially with its procedure GLM. This procedure uses the method of least squares to fit general linear models (SAS Institute [39], [40]). There are four different types of sums of squares (SS) which can be calculated in SAS. All four types can be expressed with the help of the Type I SS. The Type I SS depend on the order of the effects in the model statement. Each effect is adjusted for the preceding effects but not for the following ones. As an example regard subsequent model statement

$$\text{OUTCOME} = \text{TREATMENT CENTER TREATMENT*CENTER} \; . \tag{2.25}$$

In this case center is adjusted for treatment, while treatment is not adjusted for any other effect. The treatment-by-center interaction term is adjusted both for treatment and for center. Therefore, the sums of squares for treatment of model (2.25) will be the same as if no other effect except of the treatment effect is fitted (SAS [39])

$$\text{OUTCOME} = \text{TREATMENT} \; .$$

The remaining three types of sums of squares of SAS will be described shortly.

After the presentation of the theoretical fundamentals of multicenter trials based on fixed effects models, three particular weights w_{ct} (Table 2.1) and related terms such as treatment estimates and hypotheses (Table 2.2) will be presented in the next sections of this chapter. At the end of each section the connection of the terms with the sums of squares of the SAS procedure GLM will be described.

Table 2.1: Fixed weighting factors and derived terms

Weight	w_{ct}	$W_t = \left(\sum_c \frac{w_{ct}^2}{n_{ct}}\right)^{-1}$	$\frac{W_A W_B}{W_A+W_B}$
I	$\frac{n_{ct}}{\sum_c n_{ct}} = \frac{n_{ct}}{n_{.t}}$	$n_{.t}$	$\frac{n_{.A}\, n_{.B}}{n_{.A}+n_{.B}}$
II	$\frac{\frac{n_{cA}\, n_{cB}}{n_{cA}+n_{cB}}}{\sum_c \frac{n_{cA}\, n_{cB}}{n_{cA}+n_{cB}}}$	$\frac{\left(\sum_c \frac{n_{cA}\, n_{cB}}{n_{cA}+n_{cB}}\right)^2}{\sum_c \left(\frac{n_{cA}\, n_{cB}}{n_{cA}+n_{cB}}\right)^2 \frac{1}{n_{ct}}}$	$\sum_{c=1}^{C} \frac{n_{cA}\, n_{cB}}{n_{cA}+n_{cB}}$
III	$\frac{1}{C}$	$\frac{1}{\frac{1}{C^2}\sum_c \frac{1}{n_{ct}}}$	$\frac{C^2}{\sum_c (\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}$

Table 2.2: The different hypotheses (SAS handbook [39])

Weight	$H_0: \quad \tilde{\mu}_{.A} = \tilde{\mu}_{.B}$
I	$\frac{\sum_c n_{cA}\, \mu_{cA}}{n_{.A}} = \frac{\sum_c n_{cB}\, \mu_{cB}}{n_{.B}}$
II	$\sum_c \frac{n_{cA}\, n_{cB}}{n_{c.}}\, \mu_{cA} = \sum_c \frac{n_{cA}\, n_{cB}}{n_{c.}}\, \mu_{cB}$
III	$\sum_c \mu_{cA} = \sum_c \mu_{cB}$

2.3.1 Analysis I

If one is only interested in distinguishing between patients receiving treatment A and patients receiving treatment B, weighting factors corresponding to the sample sizes of the two treatment groups seem reasonable. In accordance with the other weighting factors which will be presented in Chapters 2.3.2 and 2.3.3, the weights will

not be assigned to the estimated treatment means $\hat{\mu}_{.A}$ and $\hat{\mu}_{.B}$ but to the estimators of the center treatment means $\hat{\mu}_{cA}$ and $\hat{\mu}_{cB}$. Thus, the quotient of the number of patients receiving treatment t at center c (n_{ct}) and the total amount of patients receiving treatment t ($n_{.t} = \sum_{c=1}^{C} n_{ct}$)

$$w_{ct}^{I} = \frac{n_{ct}}{n_{.t}} \,. \tag{2.26}$$

will be assigned to $\hat{\mu}_{ct}$. This weighting factor will be termed weight I. The BLUE for the overall mean treatment difference based on weight I is

$$\hat{\theta}^{I} \overset{(2.16)}{=} \sum_{c=1}^{C} \left(w_{cA}^{I} \, \overline{Y}_{cA.} - w_{cB}^{I} \, \overline{Y}_{cB.} \right) = \sum_{c=1}^{C} \left(\frac{n_{cA}}{n_{.A}} \, \overline{Y}_{cA.} - \frac{n_{cB}}{n_{.B}} \, \overline{Y}_{cB.} \right) \,,$$

with the following expected value and variance

$$E(\hat{\theta}^{I}) \overset{(2.17)}{=} \beta_A - \beta_B + \sum_{c=1}^{C} \left(\alpha_c \left(\frac{n_{cA}}{n_{.A}} - \frac{n_{cB}}{n_{.B}} \right) + \frac{n_{cA}}{n_{.A}} \, (\alpha\beta)_{cA} - \frac{n_{cB}}{n_{.B}} \, (\alpha\beta)_{cB} \right) \tag{2.27}$$

$$Var(\hat{\theta}^{I}) \overset{(2.18)}{=} \sigma^2 \left(\frac{W_A^I + W_B^I}{W_A^I \, W_B^I} \right) \overset{Table\ 2.1}{=} \sigma^2 \left(\frac{n_{.A} + n_{.B}}{n_{.A} \, n_{.B}} \right) \,.$$

Subsequent test statistic will result if the common variance σ^2 is not know and has to be estimated

$$\begin{aligned} T^I &= \frac{\hat{\theta}^I}{\sqrt{Var(\hat{\theta}^I)}} = \frac{\sum_{c=1}^{C} \left(\frac{n_{cA}}{n_{.A}} \, \overline{Y}_{cA.} - \frac{n_{cB}}{n_{.B}} \, \overline{Y}_{cB.} \right)}{\sqrt{\hat{\sigma}^2 \left(\frac{n_{.A}+n_{.B}}{n_{.A} \, n_{.B}} \right)}} \\ &= \frac{\frac{1}{n_{.A}} \sum_{c=1}^{C} \sum_{k=1}^{n_{cA}} Y_{cAk} - \frac{1}{n_{.B}} \sum_{c=1}^{C} \sum_{k=1}^{n_{cB}} Y_{cBk}}{\sqrt{\hat{\sigma}^2 \left(\frac{n_{.A}+n_{.B}}{n_{.A} \, n_{.B}} \right)}} \,. \end{aligned} \tag{2.28}$$

The fact that Analysis I only distinguishes between patients receiving treatment A and patients receiving treatment B is also apparent in the nominator ($\hat{\theta}^I$) of the test

statistic. $\hat{\theta}^I$ simplifies to the overall treatment estimator when the data is pooled. Hence, a one-way ANOVA with "treatment" as the only factor is obtained when weight I is applied.

Squaring the t-statistic results in the F-statistic (2.22, 2.24, Table 2.1)

$$F^I = \frac{\frac{n_{.A}\, n_{.B}}{n_{.A}+n_{.B}} \left(\tilde{Y}_{.A.} - \tilde{Y}_{.B.}\right)^2}{MSE} = \frac{\sum_{t=A}^{B} n_{.t} \left(\tilde{Y}_{.t.} - \tilde{Y}_{...}\right)^2}{MSE} = \frac{MSTreatment_w^I}{MSE} ,$$

with

$$\begin{aligned} \tilde{Y}_{.t.} &= \sum_{c=1}^{C} w_{ct}^I \overline{Y}_{ct.} = \sum_{c=1}^{C} \frac{n_{ct}}{n_{.t}} \overline{Y}_{ct.} = \frac{1}{n_{.t}} \sum_{c=1}^{C} \sum_{k=1}^{n_{ct}} Y_{ctk} \qquad \text{for} \quad t = A,\, B \\ \tilde{Y}_{...} &= \frac{\sum_{t=A}^{B} W_t^I\, \tilde{Y}_{.t.}}{\sum_{t=A}^{B} W_t^I} = \frac{\sum_{t=A}^{B} n_{.t}\, \tilde{Y}_{.t.}}{\sum_{t=A}^{B} n_{.t}} = \frac{\sum_{c=1}^{C} \sum_{t=A}^{B} \sum_{k=1}^{n_{ct}} Y_{ctk}}{n_{..}} . \end{aligned}$$

In case weighting factor I is used in (2.11), it has been proven (e.g. Neter et al. 1996 [36], Searle 1987 [44]) that the difference of the sums of squares (SSE(R)-SSE(F)) of the generalized linear model approach (2.12) simplifies to the sum of squares of a one-way ANOVA considering solely the factor "treatment", $SSTreatment_w^I = \sum_{t=A}^{B} n_{.t} \left(\tilde{Y}_{.t.} - \tilde{Y}_{...}\right)^2$.

The test statistics T^I and F^I investigate the following hypothesis ((2.23), Searle 1971 [42])

$$\begin{aligned} H_0^I : \quad & \tilde{\mu}_{.A} = \tilde{\mu}_{.B} \\ & \Longleftrightarrow \\ H_0^I : \quad & \frac{\sum_c n_{cA}\, \mu_{.A}}{n_{.A}} = \frac{\sum_c n_{cB}\, \mu_{.B}}{n_{.B}} \qquad (2.29) \\ & \Longleftrightarrow \\ H_0^I : \quad & \beta_A + \frac{1}{n_{.A}} \sum_c n_{cA} \left(\alpha_c + (\alpha\beta)_{cA}\right) = \beta_B + \frac{1}{n_{.B}} \sum_c n_{cB} \left(\alpha_c + (\alpha\beta)_{cB}\right) \end{aligned}$$

which is additionally presented in Table 2.2 for summary purposes.

With respect to the SAS procedure GLM, it has already been demonstrated that the Type I sums of squares of model

$$\text{OUTCOME} = \text{TREATMENT CENTER TREATMENT*CENTER} \quad (2.25)$$

are the same as the sums of squares for treatment of model

$$\text{OUTCOME} = \text{TREATMENT},$$

because the Type I SS depend on the ordering of the effects.

The SAS Type I SS for treatment of model (2.25) and the weighted sums of squares based on weight I correspond to the SS of a model where no other effect except the treatment effect itself had been fitted (SAS [39]). Both SS are therefore the same (Neter et al. 1996 [36], Kaufmann 2001 [30]).

The factor "treatment" is the only factor and it is not adjusted for any other effect. This fact is also apparent in the expected value of the estimator of the overall difference in treatment means (2.27). Because treatment is neither adjusted for center nor for the treatment-by-center interaction, $E(\hat{\theta}^I)$ is confounded by these two factors. Therefore, estimating the pure treatment effect $\beta_A - \beta_B$ is only possible in special cases.

2.3.2 Analysis II - Cochran, Mantel-Haenszel Estimator

The Cochran, Mantel-Haenszel estimator (CMH-estimator) of the overall treatment effect utilizes weights which are inversely proportional to the variances of the estimated treatment effects $\hat{\theta}_c$. It therefore minimizes the variance of the overall estimator $\hat{\theta} = \sum_{c=1}^{C} w_c^{CMH} \hat{\theta}_c$.

The variance of $\hat{\theta}_c$ has already shown to be (2.14)

$$Var(\hat{\theta}_c) = \sigma^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}) = v_c^2 \qquad c = 1, \ldots, C\ .$$

The weighting factor for center c (c=1,...,C) is then given by

$$w_c = \frac{\frac{1}{Var(\hat{\theta}_c)}}{\sum_{c=1}^{C} \frac{1}{Var(\hat{\theta}_c)}} = \frac{\frac{1}{\sigma^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}}{\sum_{c=1}^{C} \frac{1}{\sigma^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}}$$

$$= \frac{\frac{1}{(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}}{\sum \frac{1}{(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}} = \frac{\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}}{\sum_{c=1}^{C} \frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}} \quad \forall\, c = 1, \ldots, C \tag{2.30}$$

which will be termed weight II (Table 2.1). Although the center variances σ_c^2 are not always equal, weight II and hence an equal error variance σ^2 is often applied nonetheless. This is done primarily when the center variances σ_c^2 are assumed to be similar. A second reason for the application of equal variances are large sample sizes. Weighting factor II

$$w_c^{II} = \frac{\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}}{\sum_{c=1}^{C} \frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}}$$

is then considered to be a good approximation of

$$w_c = \frac{\frac{1}{\sigma_c^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}}{\sum_{c=1}^{C} \frac{1}{\sigma_c^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})}}.$$

As a third advantage, the application of weight II removes the problem of estimating the center variances σ_c^2, c=1,...,C which are often not known beforehand.

The differences in mean responses of the participating centers are therefore weighted depending on the total amount of patients recruited at a clinic as well as depending on the distribution of the patients to the treatment groups. A very small weighting factor will be obtained, if a center recruits only few patients and if the number of patients varies greatly between the treatment groups. In contrast, the weighting factor will get largest if many patients are enrolled at a center and if both treatment

groups have the same number of patients. In the papers of Yates (1934 [54]) and Lin (1999 [34]), an analysis is generally denoted as *weighted* when w_c^{II} is employed.

As opposed to weighting factor I, weight II does not vary with regard to the treatment levels A or B. Hence, $w_{cA}^{II} = w_{cB}^{II} = w_c^{II}$. This is a necessary requirement because the factor "treatment" is not the only factor in the model statement for analysis II. It will later be demonstrated that the model statement of analysis II contains both the factor "treatment" and the factor "center". Hence, a weighting factor that does only vary with regard to the centers is required. Otherwise the treatment effect would be confounded with the center main effects (Gallo 2000 [14]).

Applying the new weighting factor, the estimator of the overall mean treatment difference (2.16) is

$$\hat{\theta}^{II} = \sum_{c=1}^{C} w_c^{II}\,\hat{\theta}_c = \frac{1}{\left(\sum\limits_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}\right)} \sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}} * (\,\overline{Y}_{cA.} - \overline{Y}_{cB.}\,)\,,$$

with expected value and variance

$$\begin{aligned} E(\hat{\theta}^{II}) &\overset{(2.17)}{=} \beta_A - \beta_B + \sum_{c=1}^{C} w_c^{II}\,(\,\alpha\beta_{cA} - \alpha\beta_{cB}\,) \\ &= \beta_A - \beta_B + \frac{1}{\left(\sum\limits_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}\right)} \sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}} (\,(\alpha\beta_{cA}) - (\alpha\beta_{cB})\,) \quad (2.31) \end{aligned}$$

$$\begin{aligned} Var(\hat{\theta}^{II}) &\overset{(2.18)}{=} \sigma^2 \sum_{c=1}^{C} (w_c^{II})^2 \left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) \\ &= \sigma^2 \sum_{c=1}^{C} \left(\frac{\frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}}{\sum\limits_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}}\right)^2 \left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) \\ &= \sigma^2 \sum_{c=1}^{C} \frac{\frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}}{\left(\sum\limits_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}\right)^2} = \frac{\sigma^2}{\sum\limits_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}}. \end{aligned}$$

If the t-statistic

$$T^{II} = \frac{\hat{\theta}^{II}}{\sqrt{Var(\hat{\theta}^{II})}} = \frac{\frac{1}{\left(\sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}\right)} \sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}} (\bar{Y}_{cA.} - \bar{Y}_{cB.})}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}}}}$$

$$= \frac{\sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}} (\bar{Y}_{cA.} - \bar{Y}_{cB.})}{\hat{\sigma}\sqrt{\sum_{c=1}^{C} \frac{n_{cA} * n_{cB}}{n_{cA}+n_{cB}}}} . \tag{2.32}$$

is squared, the weighted F statistic $F^{II} \overset{(2.22)}{=} \frac{\sum_{t=A}^{B} W_t^{II} (\bar{Y}_{.t.} - \bar{Y}_{...})^2}{\hat{\sigma}^2}$ with W_t^{II} according to Table 2.1 will result. It examines the hypothesis

$$\begin{aligned} H_0^{II}: \quad \sum_c w_c^{II} \mu_{cA} &= \sum_c w_c^{II} \mu_{cB} \\ &\iff \\ \sum_c \frac{n_{cA} n_{cB}}{n_{c.}} \mu_{cA} &= \sum_c \frac{n_{cA} n_{cB}}{n_{c.}} \mu_{cB} \end{aligned} \tag{2.33}$$

where $n_{c.} = \sum_{t=A}^{B} n_{ct} = n_{cA} + n_{cB}$ (Table 2.2). This hypothesis is equivalent to the following hypotheses of the SAS handbook ([39] "Table 4.1: Interpretation of sums of squares in μ-model notation", page 160)

$$\begin{aligned} (1) \quad \sum_c n_{cA} \mu_{cA} &= \sum_{c=1}^{C} \sum_{t=A}^{B} \frac{n_{cA} n_{ct} \mu_{ct}}{n_{c.}} \quad \text{and} \\ (2) \quad \sum_c n_{cB} \mu_{cB} &= \sum_{c=1}^{C} \sum_{t=A}^{B} \frac{n_{cB} n_{ct} \mu_{ct}}{n_{c.}} , \end{aligned}$$

$$\text{since (1)} \iff$$

$$\sum_c n_{cA} \mu_{cA} = \sum_{c=1}^{C} \frac{n_{cA} n_{cA} \mu_{cA}}{n_{c.}} + \sum_{c=1}^{C} \frac{n_{cA} n_{cB} \mu_{cB}}{n_{c.}}$$

$$\sum_c \left(n_{cA}\,\mu_{cA} - \frac{n_{cA}\,n_{cA}\,\mu_{cA}}{n_{cA}+n_{cB}} \right) = \sum_{c=1}^{C} \frac{n_{cA}\,n_{cB}\,\mu_{cB}}{n_{c.}}$$

$$\sum_{c=1}^{C} \frac{n_{cA}\,n_{cB}\,\mu_{cA}}{n_{c.}} = \sum_{c=1}^{C} \frac{n_{cA}\,n_{cB}\,\mu_{cB}}{n_{c.}} .$$

The same equation is obtained for expression (2).

Concerning the Type II sums of squares of PROC GLM, the order of the effects in the model statement is not of importance any more. The effect under investigation, i.e. "treatment", is adjusted for all other effects in the model except any interactive effect in which treatment is involved and except factors nested within treatment (SAS handbook [39], Senn 1997 [46], Kaufmann [30]). For the underlying model (2.3), the Type II SS for treatment are equivalent to the following Type I SS for treatment

OUTCOME = CENTER TREATMENT TREATMENT*CENTER .

Treatment is therefore adjusted for the center main effect but not for the interaction. This also means that the Type II SS for treatment are the same SS of a model where the interaction effect has not been fitted at all. (Senn 1997 [46]). The Type II SS of SAS therefore correspond to Analysis II.

Searle (1971 [42]) also proved that the difference in sums of squares SSE(R)-SSE(F) of the generalized linear model approach (2.12) is in case of weighting factor II the same as the SS of a two-way anova (without interaction) where treatment is adjusted for center.

The fact that the treatment-by-center interaction is not included in the model is also apparent in the expected value of the estimator of the overall difference in treatment means (2.31). Because treatment is adjusted for center but not for the treatment-by-center interaction, $E(\hat{\theta}^{II})$ is confounded by the interaction. Estimating the pure treatment effect is also only possible in special cases.

2.3.3 Analysis III - Yates Approach

When there is no reason to consider one center as more important than any other, the analysis based on weight III, also known as *unweighted analysis or Yates Approach*, is used (Goldberg and Koury [15]). It weights all centers equally

$$w_c^\star = 1\,,$$

which means that it is independent of the number of patients at a center. Since the sum of the weights has to add up to 1 ($\sum_{c=1}^{C} w_c \stackrel{!}{=} 1$), the standardized weights are obtained as

$$w_c^{III} = \frac{w_c^\star}{\sum_{c=1}^{C} w_c^\star} = \frac{1}{C}\,.$$

They will be denoted as weights III (Table 2.1). The estimator of the overall population treatment mean difference in case of weight III is

$$\hat{\theta}^{III} = \frac{1}{C}\sum_{c=1}^{C}(\bar{Y}_{cA.} - \bar{Y}_{cB.})\,. \tag{2.34}$$

Due to the restrictions of model (2.3)

$$\sum_{c=1}^{C}(\alpha\beta)_{ct} = 0 \qquad \forall\, t = A,\, B\,,$$

the expected value of $\hat{\theta}^{III}$ is given by

$$\begin{aligned} E(\hat{\theta}^{III}) &\overset{(2.17)}{=} \frac{1}{C}\sum_{c=1}^{C}(E(\bar{Y}_{cA.}) - E(\bar{Y}_{cB.})) \\ &= \frac{1}{C}\sum_{c=1}^{C}(\mu_{cA} - \mu_{cB}) \\ &= \frac{1}{C}\sum_{c=1}^{C}(\mu_{..} + \alpha_c + \beta_A + \alpha\beta_{cA} - (\mu_{..} + \alpha_c + \beta_B + \alpha\beta_{cB})) \end{aligned}$$

$$= \beta_A - \beta_B + \frac{1}{C}\sum_{c=1}^{C}(\alpha\beta_{cA} - \alpha\beta_{cB})$$

$$= \beta_A - \beta_B \ . \tag{2.35}$$

Hence, $\hat{\theta}^{III}$ is an unbiased estimator for the overall treatment difference whether interactions are present or not. The factors β_t as well as the linear function $\beta_A - \beta_B$ are now estimable with the BLUE of $\beta_A - \beta_B$ being

$$\hat{\beta}_A - \hat{\beta}_B = \frac{1}{C}\sum_{c=1}^{C}(\overline{Y}_{cA.} - \overline{Y}_{cB.}) \, .$$

Concerning the expected values of the estimators for the overall mean treatment differences of Analysis I, II and III (formulas (2.27), (2.31), (2.35)), Analysis III is the only approach which results in an unbiased estimator. It is for this reason that in case of interactions, the Type III analysis was in former times the preferred approach.

It can again be shown that the difference in sum of squares (SSE(R)-SSE(F)) of formula (2.12) will reduce to the weighted sums of squares of treatment

$$SSTreatment_W^{III} = \sum_{t=A}^{B} \frac{C^2}{\sum_c \frac{1}{n_{ct}}} \left(\tilde{Y}_{.t.} - \tilde{Y}_{...}\right)^2$$

if weight III is employed (Searle 1987 [44]). In other words, F^{III} is the special case of the general statistic (2.12) for $w_{cA} = w_{cB} = w_c^{III}$.

The analysis based on weight III corresponds to the Type III sum of squares of SAS because in both cases the effect under investigation, i.e. treatment, is adjusted for all other effects in the model (also for the interaction) (SAS handbook [39], Senn 1997 [46], Kaufmann 2001 [30]). The order in which the effects are presented, is also not of importance for the Type III sums of squares of SAS. They are expressed with the following Type I SS for treatment

OUTCOME = CENTER TREATMENT*CENTER TREATMENT .

As mentioned earlier, there are four different types of sums of squares in SAS. The first three of them have been described so far. Type IV sums of squares are primarily useful for empty cells. In situations when all cells are filled they are the same as the Type III SS (SAS [39]). Since empty cells will not be considered in this writing there is no need for describing the Type IV SS in more detail.

2.3.4 No interactions

If an interaction between treatment and center does not exist, the difference in treatment means will be the same across all centers

$$\theta_c = \mu_{cA} - \mu_{cB} = const \quad \forall\ c = 1, ..., C\,.$$

The analyses based on weights II and III then result in the same overall difference in treatment means

$$\theta \;=\; \sum_{c=1}^{C} w_c\,(\mu_{cA} - \mu_{cB}) = const\,.$$

In addition, both analyses are unbiased if there are no interactions. The variance of the Type II analysis is however smaller since it has been proven that

$$\begin{aligned} Var(\hat{\theta}^{III}) &\geq Var(\hat{\theta}^{II}) \\ \frac{\sigma^2}{C^2}\sum_{c=1}^{C}\left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) &\geq \frac{\sigma^2}{\sum_{c=1}^{C}\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}}\,. \end{aligned}$$

In fact, $\hat{\theta}^{II}$ is the uniformly minimum variance unbiased estimator (UMVUE) of θ in case there are no interactions (Lin 1999 [34], Gallo 2000 [14]).

Applying the SAS procedure GLM it is worth mentioning that if the interaction term is left in the ANOVA model (which corresponds to the Type III analysis), although there is none or only a negligible treatment-by-center interaction, this unnecessary term will not lead to bias in the least squares estimators but it will increase their variance (Goldberg and Koury 1990 [15]).

2.3.5 Equal number of patients receiving the two treatments at each center

In this section it is assumed that an equal number of patients receives treatments A and B at center c, so that

$$n_{cA} \;=\; n_{cB} = n_c\,, \quad \forall\, c = 1, \ldots, C\,.$$

Table 2.3: Fixed weighting factors with unequal and equal sample sizes

Weight	w_{ct}		
	all n_{ct} may be different	$n_{ct} = n_c \quad \forall t$	$n_{ct} = n \quad \forall t\ \forall c$
I	$\frac{n_{ct}}{\sum_c n_{ct}} = \frac{n_{ct}}{n_{.t}}$	$\frac{n_c}{\sum_c n_c}$	$\frac{n}{Cn} = \frac{1}{C}$
II	$\frac{\frac{n_{cA}\,n_{cB}}{n_{cA}+n_{cB}}}{\sum_c \frac{n_{cA}\,n_{cB}}{n_{cA}+n_{cB}}}$	$\frac{\frac{n_c}{2}}{\sum_c \frac{n_c}{2}} = \frac{n_c}{\sum_c n_c}$	$\frac{n}{Cn} = \frac{1}{C}$
III	$\frac{1}{C}$	$\frac{1}{C}$	$\frac{1}{C}$

Table 2.3 shows in column 2 under the headline "all n_{ct} may be different" the original weighting factors. In case the number of observations for treatments A and B are the same at each center but may vary across centers, the weighting factors simplify

to the ones of column 3. Weighting factors I and II now coincide ($\frac{n_c}{\sum_c n_c}$), whereas weighting factor III remains different ($\frac{1}{C}$). Hence, in case of an equal number of patients in both treatment groups there is no need to distinguish between a model where treatment is the only factor and a model where treatment is adjusted for center.

If an equal number of patients receives treatments A and B and if additionally this number is the same at each center

$$n_{ct} \equiv n \quad \forall\, c = 1, \ldots, C \quad \forall\, t = A, B\,,$$

the three weighting factors will coincide. This is shown in the last column of Table 2.3. Analyses I to III are therefore the same.

2.4 Comparison of Analyses I to III

All three analyses will now be discussed and compared. The opinion of the "Statistical Analysis System" will be presented first, followed by the controversial point of views of different authors.

The SAS handbook (1991 [39]) states with respect to the application of the *Type I SS* that they "may not be particularly useful for the analysis of unbalanced multiway structures but may be useful for nested models, polynomial models, and certain tests involving homogeneity of regression coefficients". Hence, the Type I analysis is not regarded as being of particular importance for the underlying problem. For the application of the *Type II SS*, SAS cites situations in which there are no interactions between factor main effects. Finally, SAS recommends the *Type III SS* in situations where main effects have to be examined although interaction is present.

This is the argumentation of the "Statistical Analysis System". With regard to Analyses II and III different authors seem to have different opinions. In fact, the competing Analyses II and III have been the basis of various debate and there is no clear consensus. Some contrary point of views will now be presented.

Källén (1997 [31]) argues from a very general perspective. In his opinion one has to distinguish between two different situations. The first situation will be present if one is interested in investigating the centers - "to demonstrate that the response is the same at a number of different locations". In this case, the centers are studied. According to Källén, it is therefore natural to get the treatment difference by weighting the centers equally. On the other hand, if a center cannot provide enough patients, the second situation is present and patients are studied instead of centers. The estimator of the overall treatment effect should then be obtained in weighting the estimated center treatment effects by the size of the center. In reality however, both situations are often present. One center may not be able to provide enough patients but interest is also in demonstrating that the outcome is the same at different sites. The suggested proceeding might therefore not be of great help.

Concerning Analysis III, it has not only be seen by SAS but also by many statisticians (e.g. Yates 1934 [54], Fleiss 1986 [11], Fleiss 1985 [10]) and the pharmaceutical industry (Lin 1999 [34]) as the preferred approach when treatment-by-center interactions are present. The reason for it is the fact that Analysis III leads to an unbiased treatment estimate in contrast to Analysis II. The interactions also make the interpretation of the Type II weights almost impossible (Goldberg and Koury 1990 [15]).

A disadvantage of Analysis III is its power which is often low in comparison to Analysis II. Furthermore Analysis III can provide imprecise estimates when total sample sizes vary greatly among centers (Lin 1999 [34]). A mixture of high-order and inferior estimates then also takes place. On the other hand, Analysis II yields better power results but the above mentioned biased treatment estimate in the case of interactions. According to Neter et al. (1996 [36]) a bias should however be accepted if weight II or any other weight does reflect appropriate importance of the centers. When applying weighting factor II one therefore has to be cautious that sample sizes are not chosen arbitrarily, that they are not influenced by cost considerations or that they do not reflect unequal attrition losses in a trial (Neter et al. 1996 [36]).

If there are treatment-by-center interactions, two different kinds will generally be distinguished: quantitative and qualitative interactions (Gail and Simon 1985 [13]). If the treatment differences vary across the centers, but if they have the same direction, the interaction will be called quantitative. When the treatment differences have different directions, the interaction will be qualitative. In the case of quantitative interactions, Goldberg and Koury (1990 [15]) consider Analysis II as possible. On the other hand, when qualitative interactions exist, they prefer to describe the different interactions and to distinguish between centers where treatment A is better than treatment B, as opposed to centers where the treatment effects have the opposite direction. Although an overall treatment effect is - according to Goldberg and Koury - still well defined with qualitative interactions, applying Analysis II is

difficult in this case.

In the paper of Jones et al. (1998 [29]) a further alternative next to the usual decision making is given. The authors describe the possibility of carrying out a preliminary test (a so called pre-test) for a significant treatment-by-center interaction. The Type III estimator will then be used in case there is significant heterogeneity, while the Type II estimator will be proposed if the hypothesis of homogeneity has to be accepted. A significance level of 10% is suggested for the pre-test. However, the test has low power. It has further been shown that using the test in a similar way in case of a two-treatment two-period cross-over trial (to investigate a possible carryover effect) leads to an inflating probability of Type I error and a biased treatment estimate (Freeman 1989 [12], Jones et al. 1998 [29]). According to Freeman actual significances are higher than the nominal levels because the pre-test is highly correlated with analysis of data from the first period only. Furthermore, an unbiased treatment estimate can only be achieved if the analysis is restricted to data of the first treatment period or if it is assumed that there is no carryover effect. The investigations of Jones et al. did therefore provide only little evidence to support pre-testing.

Horwitz et al. (1996 [26]) turn away from the discussion about Analyses II and III. They mention the merits of estimates of *overall* treatment effects and their use for the pharmaceutical industry. However, a general statement on the effectiveness of a therapy tested in a heterogeneous group of patients is according to their view often not sufficient for the usefulness of the therapy for a specific patient. In this case the treating physician is generally doing a subgroup analysis. This is the reason why they divided the participating centers of the well known Beta-Blocker Heart Attack Trial (1982 [3]) into 2 groups. Group one consisted of centers where the active treatment was superior to placebo - called a dominant group. Centers with a better placebo therapy were assigned to the second group - called divergent group. After applying the Gail and Simon test (1985 [13]), they found a significant difference between the responses of the dominant and the divergent group. They concluded

their article by suggesting the search for divergent centers in multicenter trials. If the differing results of the divergent and dominant centers can not be explained by chance, the recommendations should also reflect these findings. However, Senn and Harrell (1997 [47]) are so confident that they predict "that a very similar analysis applied to any multicenter trial whatsoever, in any condition with any treatment, will always be significant at the 5% level provided only that the number of centers is at least equal to 8".

As mentioned before, Analysis II is usually regarded as being optimal when no treatment-by-center interactions are present (Fleiss [10] [11]). An article about multicenter studies considering this topic was recently published by Lin (1999 [34]). The author recommends to apply the Type II analysis not only in the absence of treatment-by-center interaction but also when treatment-by-center interaction is present. Lin considers the Type II analysis to be generally better than the Type III analysis. He basically founded his statement on the following: When treatment-by-center interaction is not present, Analysis II is the perfect match and it is additionally more powerful. On the other hand, if treatment-by-center interaction is present, it is stated in the above mentioned papers of Fleiss that the weights of Analysis II are no longer interpretable and that there is no well defined population as the basis for Analysis II. Lin now defines a study represented population (SRP). It is the "patient population represented by the multicenter study which is specific to the centers in the study and to the proportions of the patients recruited from each of the centers". In other words, "it is the unlimited patient population if the study continues forever without changing the study settings (including centres)". In the case of treatment-by-center interactions, Analysis II (more specifically $\hat{\theta}^{II}$) estimates the mean treatment difference in SRP

$$\theta_{SRP} = \sum \mathrm{P}(\text{patient belongs to center c}) * \theta_c \,.$$

It is possible to show that $\hat{\theta}^{II}$ is asymptotically the UMVUE of θ_{SRP}. In contrast, the unweighted analysis ($\hat{\theta}^{III}$) estimates the average treatment difference over all

centers ($\theta^{III} = \frac{1}{C}\sum\theta_c = \theta_{AVE}$). The question is now which of the overall mean treatment differences (θ_{SRP} or θ_{AVE}) is the closer approximation of the mean treatment difference in the targeted patient population as defined in the study protocol (θ_T). Lin further argues that this is dependent on the centers. He states that in reality, different centers usually serve different sizes of patient population and the centers may also not be equally motivated in recruiting patients. Considering the risk of losing power with Analysis III, he suggests to prefer Analysis II in general. Furthermore, even if Analysis III is the closer approximation of the mean treatment difference in the targeted population (θ_T), Lin will recommend to use Analysis II in the case of greatly varying sample sizes between centers.

Overall, the Type II analysis should be preferred if there are no interactions. When treatment-by-center interactions are present, a first attempt should of course be made in finding center specific attributes that affect the treatments (Gould 1998 [16]). If no covariates for the explanation of the interactions can be detected, the Type II analysis should be applied if the sample sizes vary greatly between the centers or if weight II does reflect the appropriate importance of the centers. The latter point gives of course way to discussion and interpretation. If the Type III analysis *must* be used for any reason when the sample sizes are unequal, the conventionally calculated sample size should be increased according to suggestions made by Gallo (2000 [14]). This augmentation compensates for the loss in power.

However, there are many situations where equal reasons for the application of the Type II and III analyses exist. The only solution seems to be the calculation of both and their comparison afterwards.

Before this is further investigated within the framework of several simulation studies, a completely different weighting factor will be presented in the next chapter. It is based on a model with a random center effect and a random treatment-by-center interaction.

Chapter 3

Random center effects

Until now, center and treatment effects as well as the treatment-by-center interaction have been chosen as fixed effects in the ANOVA model. This was done since interest was in the particular centers themselves as well as in the two specific treatments. The factor "treatment" will remain fixed. However, the factor "center" can also be regarded as random. The participating centers are then considered as a randomly chosen sample from a population of centers. Hence, the individual clinics are of no interest of themselves but represent the population they have been selected from. Inferences drawn from the random sample of clinics can be generalized to the whole population.

Although models with random effects have been described in the literature, they do not obtain the same attention as fixed effects models - at least with regard to multicenter clinical trials. This is probably due to the fact that neither a fixed nor a random center effect seem to be a perfect fit (Fleiss 1986 [11]). Since it is not easy to describe the population of which the random sample of centers is selected from, a fixed center effect is then often preferred. However, the use of random effects has recently been enforced as a means of taking heterogeneity of treatment effects into account (Hardy and Thompson 1996 [18]). According to Gould (1998 [16]) "this can be a convenient approximation even though it is certainly true that centers are not

randomly sampled from some plausible population of centers". Finally, concerning the recommendation of the International Conference on Harmonization (ICH), their E9 guideline terms a random center effect as particularly relevant when the number of sites is large.

As described in Chapters 1 and 2, the complexity of fixed effect models increases considerably in turning away from the balanced case to unbalanced data. According to Neter et al. (1996 [36]) the level of complexity increases in a similar way for studies involving random effects when sample sizes are unequal.

With regard to the terminology of ANOVA models, the underlying model of a two-factor analysis of variance with a fixed treatment effect, a random center effect as well as a random treatment-by-center interaction is a **mixed effects model**. ANOVA models will in general be denoted as mixed, if at least one fixed and one random factor are present. Models consisting of fixed effects only are called **fixed effect models**, whereas in case of random effects only, the models are denoted as **random**.

3.1 ANOVA Model

The factor effects mixed ANOVA model with a fixed treatment effect, a random center effect as well as a random treatment-by-center interaction is similar to model (2.3)

$$Y_{ctk} = \mu_{..} + \alpha_c + \beta_t + (\alpha\beta)_{ct} + e_{ctk} \ , \tag{3.1}$$

however with (Neter et al. 1996 [36])

- the general mean $\mu_{..}$ being a constant,
- center main effects α_c which are random variables, independent and identically $N(0, \sigma_\alpha^2)$ distributed,

- treatment main effects β_t which are constants subject to the restriction $\sum_{t=A}^{B} \beta_t = 0$. The number of treatment levels is T=2, because t can be either A or B.

- The treatment-by-center interaction effects $(\alpha\beta)_{ct}$ are $N(0, \frac{T-1}{T}\,\sigma^2_{\alpha\beta}) = N(0, \frac{1}{2}\,\sigma^2_{\alpha\beta})$ distributed random variables subject to the restrictions

  $$\sum_{t=A}^{B} (\alpha\beta)_{ct} = 0, \quad \forall\, c = 1, \ldots, C\,.$$

  The reason why $\frac{1}{2}\,\sigma^2_{\alpha\beta}$ and not $\sigma^2_{\alpha\beta}$ is used as variance of the interaction term is that the expected mean squares will then be simpler expressions. Any two interaction terms are independent except if they come from the same center. In this case they will be correlated with the covariance being

  $$Cov[(\alpha\beta)_{cA}, (\alpha\beta)_{cB}] = -\tfrac{1}{T}\,\sigma^2_{\alpha\beta} = -\tfrac{1}{2}\,\sigma^2_{\alpha\beta}\,.$$

- The random error terms e_{ctk} are independent $N(0, \sigma^2_e)$ distributed.

- The α_c, $(\alpha\beta)_{ct}$, and e_{ctk} are pairwise independent,

- c=1, ... , C; t=A, B; k=1, ... , n_{ct} .

In a fixed effects model, the random error terms are defined as $e_{ctk} = Y_{ctk} - E[Y_{ctk}]$. Now, in order to take the random center effects into account the error terms are differently defined as

$$e_{ctk} = Y_{ctk} - E[\,Y_{ctk} \mid \alpha_c,\, (\alpha\beta)_{ct}\,] = Y_{ctk} - (\mu_{..} + \alpha_c + \beta_t + (\alpha\beta)_{ct})\ ,$$

where $E[\,Y_{ctk} \mid \alpha_c,\, (\alpha\beta)_{ct}\,]$ is the expected value of Y_{ctk} for given α_c and $(\alpha\beta)_{ct}$ (Searle 1987 [44]).

The expected value and the variance for the responses of this model are given by (Neter et al. 1996 [36])

$$\begin{aligned} E[Y_{ctk}] &= \mu_{..} + \beta_t \qquad \text{since} \quad E(\alpha_c) = 0,\ E((\alpha\beta)_{ct}) = 0 \\ Var[Y_{ctk}] &= \sigma^2_Y = \sigma^2_\alpha + \frac{T-1}{T}\,\sigma^2_{\alpha\beta} + \sigma^2_e = \sigma^2_\alpha + \frac{\sigma^2_{\alpha\beta}}{2} + \sigma^2_e\,. \end{aligned} \tag{3.2}$$

As can be seen from formula (3.2), the variance of the response Y_{ctk} of a mixed model does not only depend on the error variance σ_e^2 but also on other variances. The individual variances are denoted as **variance components**. In the fixed model analysis, the only variance σ_e^2 does not influence the estimation of the fixed effects. The problem with mixed effects models is that the variance components do affect the estimation of the fixed effects. Furthermore, their estimation is often difficult. This is also known as the **mixed model problem** (Searle 1987 [44]).

The observations Y_{ctk} will now only be independent if they belong to different levels of the factor center and hence if they belong to different centers. This is in contrast to the fixed effects model where they are all independent. In case the responses are from the same center, they are correlated with the following covariances

$$\begin{aligned}
\sigma(Y_{ctk}, Y_{ctk'}) &= \sigma_\alpha^2 + \frac{T-1}{T}\,\sigma_{\alpha\beta}^2 = \sigma_\alpha^2 + \frac{1}{2}\,\sigma_{\alpha\beta}^2\,, \qquad k \neq k' \\
\sigma(Y_{cAk}, Y_{cBk'}) &= \sigma_\alpha^2 - \frac{1}{T}\,\sigma_{\alpha\beta}^2 = \sigma_\alpha^2 - \frac{1}{2}\,\sigma_{\alpha\beta}^2\,, \qquad A \neq B \\
\sigma(Y_{cAk}, Y_{c'Bk'}) &= 0\,, \qquad c \neq c'\,.
\end{aligned}$$

3.2 General Mixed Model

Mixed model (3.1) can also be expressed with the matrix notation of a general mixed model as described in Searle (1992 [45]).

Denote $\underline{\mu}$ as the cell means vector of all fixed effects, $\underline{u}_1$ as the vector of the random center effects and $\underline{u}_2$ as the vector of the random treatment-by-center interaction terms. $\underline{\underline{X}}$, $\underline{\underline{Z}}_1$ and $\underline{\underline{Z}}_2$ are the corresponding design matrices. Model (3.1) can then be rewritten in a more general way as

$$\underline{Y} = \underline{\underline{X}}\,\underline{\mu} + \underline{\underline{Z}}_1\,\underline{u}_1 + \underline{\underline{Z}}_2\,\underline{u}_2 + \underline{e} = \underline{\underline{X}}\,\underline{\mu} + \sum_{i=1}^{2} \underline{\underline{Z}}_i\,\underline{u}_i + \underline{e}\,.$$

The expected values of all random effects are zero

$$E(\underline{u}_1) = 0\,, \qquad E(\underline{u}_2) = 0\,, \qquad E(\underline{e}) = 0\,,$$

leading to

$$E(\underline{Y}) \;=\; \underline{\underline{X}}\,\underline{\mu}\,.$$

The expected value of the response vector $\underline{Y}$ conditional on the random effects enclosed in $\underline{u}_1$ and $\underline{u}_2$ is given by

$$E(\underline{Y} \mid \underline{u}_1, \underline{u}_2) \;=\; \underline{\underline{X}}\,\underline{\mu} + \underline{\underline{Z}}_1\,\underline{u}_1 + \underline{\underline{Z}}_2\,\underline{u}_2\,.$$

Finally, the variance of the response vector $\underline{Y}$ expressed in the general form is

$$\begin{aligned} Var(\underline{Y}) \;&=\; \underline{\underline{V}} = \sigma_Y^2 = Var\left(\sum_{i=1}^{2} \underline{\underline{Z}}_i\,\underline{u}_i + \underline{e}\right) \\ &=\; \sum_{i=1}^{2} \underline{\underline{Z}}_i\,\underline{\underline{Z}}_i^T\,\sigma_i^2 + \sigma_e^2\,\underline{\underline{I}}_{n_{..}} \end{aligned} \tag{3.3}$$

with

$$\begin{aligned} \underline{\underline{I}}_{n_{..}} \qquad & \text{representing the } n_{..} \times n_{..} \text{ identity matrix} \\ Var(\underline{u}_i) \;&=\; \sigma_i^2\,\underline{\underline{I}}_{q_i} \quad \forall\, i = 1, 2; \quad q_i \text{ is the order of } \underline{u}_i \\ Var(\underline{e}) \;&=\; \sigma_e^2\,\underline{\underline{I}}_{n_{..}} \\ Cov(\underline{u}_1, \underline{u}_2^T) \;&=\; 0 \\ Cov(\underline{u}_i, \underline{e}^T) \;&=\; 0 \quad \forall\, i = 1, 2\,. \end{aligned}$$

If the design matrix $\underline{\underline{X}}$ has full column rank, the parameter vector $\underline{\mu}$ will be estimable applying **weighted least squares** (a special case of generalized least squares). It requires the minimization of

$$(\underline{y} - \underline{\underline{X}}\,\underline{\mu})^T\,\underline{\underline{V}}^{-1}\,(\underline{y} - \underline{\underline{X}}\,\underline{\mu})\,,$$

with $\underline{\underline{V}}$ being covariance matrix (3.3). The resulting **mixed model equations** can be solved for $\hat{\underline{\mu}}$ and $\hat{\underline{u}}$. The BLUE of $\underline{\mu}$ then is

$$\hat{\underline{\mu}} = \left(\underline{\underline{X}}^T \, \underline{\underline{V}}^{-1} \, \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \, \underline{\underline{V}}^{-1} \, \underline{y} \,, \tag{3.4}$$

where the variance of the estimator of the parameter vector $\underline{\mu}$ is given by

$$Var(\hat{\underline{\mu}}) = \left(\underline{\underline{X}}^T \, \underline{\underline{V}}^{-1} \, \underline{\underline{X}}\right)^{-1} .$$

The best linear unbiased predictor[1] of the vector of all random components $\underline{u}$ is defined as

$$\hat{\underline{u}} = \underline{\underline{G}}_u \, \underline{\underline{Z}}^T \, \underline{\underline{V}}^{-1} \left(\underline{y} - \underline{\underline{X}} \, \hat{\underline{\mu}}\right) .$$

The matrix $\underline{\underline{G}}_u$ contains the variance components σ_i^2, $i = 1, 2$, in a diagonal structure. $\underline{\underline{Z}}$ is the design matrix for the vector $\underline{u}$ of all random components. The difference to the fixed effects case - where the parameters had been estimated with the help of ordinary least squares (2.6) - is the covariance matrix $\underline{\underline{V}}$. Utilizing $\hat{\underline{\mu}} = \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \underline{y}$ for a mixed model and therefore not considering the variance components still results in parameter estimates but not always in BLU estimators (because the estimates do not necessarily have minimum variance). In case the covariance matrix is not known, its estimation $\hat{\underline{\underline{V}}}$ is commonly employed instead.

The variance components of the matrix $\underline{\underline{V}}$ can be estimated in a number of different ways. Choosing one method might not be an easy task. According to Searle (1992 [45]) there are 9 methods available: a general ANOVA method, Henderson's methods I to III (although only Henderson's method III is appropriate for a mixed model with an interaction between a fixed and a random factor), Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), the non-iterative

[1] Linear combinations of fixed *and* random effects are not called estimable functions but *predictable functions*. Solving the mixed model equations yields *predictors*, or *BLUPs* (SAS System for Mixed Models, 1996). They are conditional means of the random effects.

Minimum Norm Quadratic Unbiased Estimation (MINQUE), Minimum Variance Quadratic Unbiased Estimation (MIVQUE) and Iterative MINQUE (I-MINQUE). The ANOVA method as well as Henderson's methods (which are just special forms of the general ANOVA procedure) have the disadvantages that for unbalanced data optimal application of the methods is not possible and distributional properties of the estimators are not known (Searle 1987 [44]). Moreover, the methods can yield negative estimates.

The computational evaluation of mixed models is usually done with the help of the PROC MIXED procedure in SAS. According to SAS (1999 [40]) "the MIXED procedure is the flagship procedure for random and mixed effects linear models just as PROC GLM is the flagship procedure for fixed effects linear models". The mixed procedure is based on the likelihood function and associated theory. PROC MIXED estimates the variance components based on a full or a restricted likelihood function (ML or REML estimation). As mentioned before, the maximum likelihood estimates of the unknown parameters are those values which maximize the probability density function of the data. The difference between the ML and the REML estimation is that the likelihood to be maximized is slightly changed for REML estimation. REML estimation thereby adjusts for the fact that the parameters and variances are estimated from the same data set (DerSimonian et al. 1986 [7]). Or, in other words, "restricted maximum likelihood maximizes that part of the likelihood which is invariant to the location parameters of the model, i.e. to the fixed effects" (Searle 1992 [45]). The log likelihood functions for both approaches can be written in a way, so that they only depend on the variance components σ_i^2, σ_e^2 (SAS 1999 [40])

$$\begin{aligned} ML: \quad l(\sigma_i^2, \sigma_e^2) &= -\frac{1}{2} log\,|\underline{\underline{V}}| - \frac{n}{2} log\, \underline{\underline{r}}^T \underline{\underline{V}}^{-1} \underline{\underline{r}} - \frac{n}{2} log\,(2\pi) \\ REML: \quad l_R(\sigma_i^2, \sigma_e^2) &= -\frac{1}{2} log\,|\underline{\underline{V}}| - \frac{1}{2} log\,|\underline{\underline{X}}^T \underline{\underline{V}}^{-1} \underline{\underline{X}}| - \frac{1}{2} \underline{\underline{r}}^T \underline{\underline{V}}^{-1} \underline{\underline{r}} \\ &\qquad - \frac{n-p}{2} log\,(2\pi)\,, \end{aligned}$$

where $\underline{\underline{r}} = \underline{Y} - \underline{\underline{X}}\,(\underline{\underline{X}}^T \underline{\underline{V}}^{-1}\, \underline{\underline{X}})^- \underline{\underline{X}}^T \underline{\underline{V}}^{-1}\, \underline{y}$, p is the rank of $\underline{\underline{X}}$ and $(\underline{\underline{X}}^T \underline{\underline{V}}^{-1}\, \underline{\underline{X}})^-$

is the generalized inverse of $(\underline{\underline{X}}^T \underline{\underline{V}}^{-1} \underline{\underline{X}})$. The variance estimates are obtained in maximizing one of the log likelihood functions. This is done in applying a Newton-Raphson algorithm. The resulting variance estimates are then used for the calculation of the parameter estimates $\underline{\hat{\mu}}$ (3.4).

3.3 Estimation of overall treatment effect

As for the fixed effects model, the matrix notation (of Chapter 3.2) will also not be considered further in the mixed model case. The cell means notation, already introduced in Chapter 2.3, will be applied for mixed model (3.1) too. With respect to multicenter studies, the mixed effects model is often also denoted as the "random approach" in comparison to the "fixed approach" although the model contains both random and fixed effects.

The random approach by DerSimonian et al. (1986 [7]) will now be described. It allows the treatment effects of the centers to vary and thereby accounts for heterogeneity. Heterogeneity is defined as different treatment estimates effectuated by the centers. The approach by DerSimonian et al. assumes that the treatment effects do have a distribution. This distribution will be estimated with the help of the observed effects of the individual centers.

The observed treatment effect of each center is defined as the sum of two random variables, the true treatment effect θ_c^r and the sampling error e_c

$$\hat{\theta}_c^r \;=\; \overline{Y}_{cA.} - \overline{Y}_{cB.} = \; \theta_c^r \; + \; e_c \qquad c = 1, ..., C \;, \tag{3.5}$$

where

$$e_c = \overline{e}_{cA.} - \overline{e}_{cB.} = \frac{1}{n_{cA}} \sum_{k=1}^{n_{cA}} e_{cAk} - \frac{1}{n_{cB}} \sum_{k=1}^{n_{cB}} e_{cBk}$$

is $N(0, v_c^2) = N(0, \sigma^2 (\frac{1}{n_{cA}} + \frac{1}{n_{cB}}))$ distributed with σ^2 corresponding to the error variability σ_e^2 of model (3.1).

The true treatment effect of each center, θ_c^r, will depend on patient characteristics, design execution in each center, etc. It will therefore vary across centers. To take this variation into account, DerSimonian et al. assume that the true treatment effect is the sum of a fixed and a random variable, i.e. the mean treatment effect for a population (θ^r) plus the deviation (δ_c) of the treatment effect of the cth center from the population mean. Hence,

$$\theta_c^r \;=\; \theta^r \;+\; \delta_c \qquad \text{c=1,...,C} \; . \tag{3.6}$$

Some distributional properties of the treatment effects will be given now.

The expected value of θ_c^r is the fixed effect θ^r since the expected value of δ_c equals 0. The variance of the deviation δ_c and therefore also of the true treatment effect θ_c^r is the between-center-variability σ_B^2. According to DerSimonian et al. σ_B^2 represents both the degree to which treatment effects vary across centers as well as the degree to which individual centers give biased assessments of treatment effects. Assuming normality (which will also be presumed for the maximum likelihood estimates in 3.3.2 because it is stringent there), the true treatment effect for center c is

$$\theta_c^r \sim N(\theta^r, \sigma_B^2) \qquad \text{distributed.}$$

The connection with model (3.1) will be apparent if the true treatment effect of center c (θ_c^r) is expressed with the components of model (3.1)

$$\theta_c^r \;=\; \theta^r \;+\; \delta_c \overset{(3.1)}{=} \mu_{cA} - \mu_{cB} = \beta_A - \beta_B + (\alpha\beta)_{cA} - (\alpha\beta)_{cB} \, , \tag{3.7}$$

where $\beta_A - \beta_B$ represent the mean treatment effect θ^r. The expected value and the variance of the true treatment effect for center c are then given by

$$E[\theta_c^r] \;=\; E[\, \beta_A - \beta_B + (\alpha\beta)_{cA} - (\alpha\beta)_{cB} \,] = \beta_A - \beta_B = \theta^r \tag{3.8}$$

$$
\begin{aligned}
Var[\theta_c^r] &= Var[\,(\alpha\beta)_{cA} - (\alpha\beta)_{cB}\,] \\
&= Var[\,(\alpha\beta)_{cA}\,] + Var[\,(\alpha\beta)_{cB}\,] + 2\,(1)(-1)\,Cov[\,(\alpha\beta)_{cA}, (\alpha\beta)_{cB}\,] \\
&\overset{(3.1)}{=} \frac{1}{2}\,\sigma^2_{\alpha\beta} + \frac{1}{2}\,\sigma^2_{\alpha\beta} - 2\,(-\frac{1}{2}\sigma^2_{\alpha\beta}) = 2\,\sigma^2_{\alpha\beta} = \sigma^2_B\,, \qquad (3.9)
\end{aligned}
$$

applying the conditions of model (3.1). σ^2_B therefore equals $2\,\sigma^2_{\alpha\beta}$.

Rewriting (3.5) with the help of (3.6) yields

$$\hat{\theta}_c^r \;=\; \overline{Y}_{cA.} - \overline{Y}_{cB.} = \;\theta_c^r \;+\; e_c = \theta^r \;+\; \delta_c + e_c\,.$$

The conditional mean and variance of $\hat{\theta}_c^r$ - which correspond to the mean and variance of $\hat{\theta}_c^r$ for given deviation δ_c - are

$$
\begin{aligned}
E[\hat{\theta}_c^r \mid \delta_c] &= \theta_c^r \\
Var[\hat{\theta}_c^r \mid \delta_c] &= v_c^2\,.
\end{aligned}
$$

Under the assumption of normality, the conditional distribution of the treatment mean estimator for center c is therefore given by

$$F_{\hat{\theta}_c^r \,|\, \delta_c} = N(\theta_c^r,\, v_c^2)\,. \qquad (3.10)$$

The marginal distribution of $\hat{\theta}_c^r$ is normal with mean θ^r and variance $v_c^2 + \sigma_B^2$ (Whitehead and Whitehead 1991 [53])

$$F_{\hat{\theta}_c^r} = \int_y f_{\hat{\theta}_c^r \,|\, \theta_c^r}(x, y)\, f_{\theta_c^r}(y)\, dy = N(\theta^r,\, v_c^2 + \sigma_B^2)\,. \qquad (3.11)$$

This can also be demonstrated if the terminology of the original model (3.1) is applied since

$$E\,[\hat{\theta}_c^r] \;=\; E\,[\,\overline{Y}_{cA.} - \overline{Y}_{cB.}\,]$$

$$
\begin{aligned}
&= E[\frac{1}{n_{cA}}\sum_{k=1}^{n_{cA}} Y_{cAk} - \frac{1}{n_{cB}}\sum_{k=1}^{n_{cB}} Y_{cBk}\,] \\
&\overset{(3.1)}{=} E\,[\,\frac{1}{n_{cA}}\sum_{k=1}^{n_{cA}}(\mu_{..}+\alpha_c+\beta_A+(\alpha\beta)_{cA}+e_{cAk}) \\
&\qquad\qquad - \frac{1}{n_{cB}}\sum_{k=1}^{n_{cB}}(\mu_{..}+\alpha_c+\beta_B+(\alpha\beta)_{cB}+e_{cBk})\,] \\
&= E\,[\,\mu_{..}+\alpha_c+\beta_A+(\alpha\beta)_{cA}+\frac{1}{n_{cA}}\sum_{k=1}^{n_{cA}} e_{cAk} \\
&\qquad\qquad - (\,\mu_{..}+\alpha_c+\beta_B+(\alpha\beta)_{cB}+\frac{1}{n_{cB}}\sum_{k=1}^{n_{cB}} e_{cBk}\,)\,] \\
&= E\,[\beta_A-\beta_B+(\alpha\beta)_{cA}-(\alpha\beta)_{cB}+\frac{1}{n_{cA}}\sum_{k=1}^{n_{cA}} e_{cAk}-\frac{1}{n_{cB}}\sum_{k=1}^{n_{cB}} e_{cBk}] \\
&= \beta_A-\beta_B=\theta^r \qquad (3.12)
\end{aligned}
$$

$$
\begin{aligned}
Var\,[\hat{\theta}_c^r] &= Var\,[\beta_A-\beta_B+(\alpha\beta)_{cA}-(\alpha\beta)_{cB}+\frac{1}{n_{cA}}\sum_{k=1}^{n_{cA}} e_{cAk}-\frac{1}{n_{cB}}\sum_{k=1}^{n_{cB}} e_{cBk}] \\
&= Var[(\alpha\beta)_{cA}-(\alpha\beta)_{cB}]+Var[\frac{1}{n_{cA}}\sum_{k=1}^{n_{cA}} e_{cAk}-\frac{1}{n_{cB}}\sum_{k=1}^{n_{cB}} e_{cBk}] \\
&\overset{(3.9)}{=} 2\,\sigma^2_{\alpha\beta}+\sigma^2\,(\frac{1}{n_{cA}}+\frac{1}{n_{cB}}) = 2\,\sigma^2_{\alpha\beta}+v_c^2=\sigma_B^2+v_c^2\,. \qquad (3.13)
\end{aligned}
$$

In order to estimate the overall treatment effect θ^r of mixed model (3.1), the center treatment effects θ_c^r - and therefore the centers - are weighted by the inverse of their variance as was performed for the Type II analysis of a fixed effects model. Hence, the θ_c^r are weighted by the sum of the within-center-variance v_c^2 and the between-center-variability σ_B^2. The resulting weighting factor for center c (c=1,...,C) now looks as follows (DerSimonian et al. 1986 [7])

$$
w_c^{\star,r} = \frac{1}{\text{within + between center variance}} = \frac{1}{v_c^2+\sigma_B^2}\,. \qquad (3.14)
$$

As previously mentioned, the between-center-variance σ_B^2 reflects unexplained heterogeneity between the centers. It is assumed to be non-negative (Hartung and Knapp 2001 [21]). The proceeding of not only including the within-center-variance but also the between-center-variance into the weighting factor provides a method for weighting which can be made more general by incorporating study characteristics into the analysis (DerSimonian et al. 1986 [7]). Hence, it represents a method amenable for generalization. This will be performed in Chapter 5 based on a fixed effects model.

The overall treatment effect of mixed model (3.1) based on weight (3.14) is defined as

$$\theta^r = \frac{\sum_{c=1}^{C} w_c^{\star,r}\ \theta_c^r}{\sum_{c=1}^{C} w_c^{\star,r}}, \qquad \text{with} \quad w_c^{\star,r} = \frac{1}{v_c^2 + \sigma_B^2}. \tag{3.15}$$

Cochran's semiweighted estimator of the overall treatment effect (Cochran 1983 [6]) is given by

$$\hat{\theta}^r = \frac{\sum_{c=1}^{C} w_c^{\star,r}\ \hat{\theta}_c^r}{\sum_{c=1}^{C} w_c^{\star,r}} = \frac{\sum_{c=1}^{C} w_c^{\star,r}\ (\overline{Y}_{cA.} - \overline{Y}_{cB.})}{\sum_{c=1}^{C} w_c^{\star,r}}, \tag{3.16}$$

with expected value

$$E(\hat{\theta}^r) \;=\; E\left(\frac{\sum_{c=1}^{C} w_c^{\star,r}\ \hat{\theta}_c^r}{\sum_{c=1}^{C} w_c^{\star,r}}\right) = \frac{\sum_{c=1}^{C} w_c^{\star,r}\ E(\hat{\theta}_c^r)}{\sum_{c=1}^{C} w_c^{\star,r}} = \theta^r \tag{3.17}$$

and variance

$$Var(\hat{\theta}^r) \;=\; \frac{\sum_{c=1}^{C} (w_c^{\star,r})^2\ Var(\hat{\theta}_c^r)}{\left(\sum_{c=1}^{C} w_c^{\star,r}\right)^2}$$

$$\overset{(3.13)}{=} \frac{\sum_{c=1}^{C}(w_c^{\star,r})^2 \ (v_c^2+\sigma_B^2)}{\left(\sum_{c=1}^{C} w_c^{\star,r}\right)^2} \overset{(3.14)}{=} \frac{1}{\sum_{c=1}^{C} w_c^{\star,r}} . \tag{3.18}$$

The variances v_c^2 and σ_B^2 are assumed to be known while in practice they often have to be estimated. There are two commonly employed methods for the calculation of the estimator of the overall treatment effect θ^r as well as of the estimator of the between-center-variability σ_B^2. They will be described in the next two sections.

3.3.1 DerSimonian and Laird Method

DerSimonian and Laird (1986 [7]) propose to estimate the overall treatment effect according to (3.15)

$$\hat{\theta}^r = \frac{\sum_{c=1}^{C} w_c^{\star,r}\ \hat{\theta}_c^r}{\sum_{c=1}^{C} w_c^{\star,r}}, \qquad \text{where} \quad w_c^{\star,r} = \frac{1}{v_c^2+\hat{\sigma}_B^2} \tag{3.19}$$

$$\hat{\theta}_c^r = \overline{Y}_{cA.} - \overline{Y}_{cB.} .$$

The only difference to (3.15) is the estimator $\hat{\sigma}_B^2$ which is applied instead of σ_B^2. The variance of the estimated overall treatment effect (3.18), given by the inverse of the added weights, is then only an approximate solution. Using the inverse of the overall variance as weighting factor guarantees an estimate with smallest variance. The sum of the weights $w_c^{\star,r}$ is sometimes denoted as the total information I.

DerSimonian and Laird suggest to estimate the between-center-variance σ_B^2 with the help of Cochran's large sample statistic Q (1954 [5])

$$Q = \sum_{c=1}^{C} w_c^{\star}\,(\hat{\theta}_c^r - \hat{\theta}^r)^2 = \sum_{c=1}^{C} w_c^{\star} \left(\hat{\theta}_c^r - \frac{\sum_c w_c^{\star}\,\hat{\theta}_c^r}{\sum_c w_c^{\star}}\right)^2 \overset{H_0}{\sim} \chi^2_{C-1} , \tag{3.20}$$

which investigates heterogeneity of treatment effects. The weight $w_c^\star$ used in the latter formula is the inverse of the variance v_c^2. In case it is also unknown, it might be estimated with the help of the error mean square (2.20)

$$\hat{v}_c^2 = \hat{\sigma}^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}) \; .$$

An estimator of σ_B^2 will then be obtained by equating the large sample test statistic Q with its expectation. The **moment estimator** of σ_B^2 results

$$\hat{\sigma}_B^2 = \frac{Q - (C-1)}{\sum\limits_c w_c^\star - \frac{\sum\limits_c (w_c^\star)^2}{\sum\limits_c w_c^\star}} \; .$$

A disadvantage of the DerSimonian and Laird method (3.19) is the fact that it does not consider that σ_B^2 has to be estimated. As a result, the confidence interval for the overall treatment difference will be too narrow, although the confidence interval is already wider than the one of the fixed effects model. With a small number of participating centers, σ_B^2 will not be estimated precisely. Furthermore, it sometimes happens that the obtained estimator is negative. Since it is impossible that a variance is negative, $\hat{\sigma}_B^2$ is set to zero in this case

$$\hat{\sigma}_B^2 = max \; \{ \; 0 \, , \; \frac{Q - (C-1)}{\sum\limits_c w_c^\star - \frac{\sum\limits_c (w_c^\star)^2}{\sum\limits_c w_c^\star}} \; \} \; .$$

The truncation is responsible for the estimator $\hat{\sigma}_B^2$ being biased, with a positive probability to take on the value 0.

As a last disadvantage of the DerSimonian and Laird estimator it has to be mentioned that no confidence interval for the between-center-variance σ_B^2 is provided. However, in spite of all its disadvantages, the moment estimator is often used in practice because of its simplicity and non-iterativeness (Thompson and Sharp 1999 [51]).

Equal within center variances

If the variances v_c^2 are all equal ($v_c^2 \equiv v^2 \; \forall c$), the Type III estimator of the fixed effects model $\hat{\theta}^{III}$ (2.34) will be obtained (DerSimonian and Laird 1986 [7]). This is due to the fact that the weights simplify to

$$w_c^{\star,r} = \frac{1}{v^2 + \sigma_B^2},$$

resulting in an overall treatment estimator (3.16) of

$$\hat{\theta}^r = \frac{\sum_{c=1}^{C} w_c^{\star,r}\, \hat{\theta}_c}{\sum_{c=1}^{C} w_c^{\star,r}} = \frac{\sum_{c=1}^{C} \frac{1}{v^2+\sigma_B^2}\, \hat{\theta}_c}{\sum_{c=1}^{C} \frac{1}{v^2+\sigma_B^2}} = \frac{1}{C}\sum_{c=1}^{C}(\overline{Y}_{cA.} - \overline{Y}_{cB.}) \; .$$

The random effects estimator of the overall difference in treatment means is therefore equivalent to the fixed effects estimator when centers are assumed to have equal importance. The formulas for the statistic Q as well as for the estimator of the between-center-variance also simplify in case of equal within-center-variances (DerSimonian and Laird 1986 [7])

$$\begin{aligned}
Q &= \sum_{c=1}^{C} w_c^{\star}\,(\hat{\theta}_c^r - \hat{\theta}^r)^2 = \sum_{c=1}^{C} \frac{1}{v^2}\left((\overline{Y}_{cA.} - \overline{Y}_{cB.}) - \frac{\sum_c \frac{1}{v^2}(\overline{Y}_{cA.} - \overline{Y}_{cB.})}{\sum_c \frac{1}{v^2}}\right)^2 \\
&= \frac{1}{v^2}\sum_{c=1}^{C}\left((\overline{Y}_{cA.} - \overline{Y}_{cB.}) - \frac{\frac{1}{v^2}\sum_c(\overline{Y}_{cA.} - \overline{Y}_{cB.})}{\frac{C}{v^2}}\right)^2 \\
&= \frac{1}{v^2}\sum_{c=1}^{C}\left((\overline{Y}_{cA.} - \overline{Y}_{cB.}) - \frac{1}{C}\sum_c(\overline{Y}_{cA.} - \overline{Y}_{cB.})\right)^2 \\
&= \frac{1}{v^2}\sum_{c=1}^{C}\left(\hat{\theta}_c^r - \frac{1}{C}\sum_c \hat{\theta}_c^r\right)^2 \\
\hat{\sigma}_B^2 &= \frac{Q-(C-1)}{\sum_c w_c^{\star} - \frac{\sum_c (w_c^{\star})^2}{\sum_c w_c^{\star}}} = \frac{\frac{1}{v^2}\sum_{c=1}^{C}\left(\hat{\theta}_c^r - \frac{1}{C}\sum_c \hat{\theta}_c^r\right)^2 - (C-1)}{\sum_c \frac{1}{v^2} - \frac{\sum_c (\frac{1}{v^2})^2}{\sum_c \frac{1}{v^2}}}
\end{aligned}$$

$$= \frac{\frac{1}{v^2}\sum_{c=1}^{C}\left(\hat{\theta}_c^r - \frac{1}{C}\sum_c \hat{\theta}_c^r\right)^2 - (C-1)}{\frac{1}{v^2}(C-1)}$$

$$= \frac{\sum_{c=1}^{C}\left(\hat{\theta}_c^r - \frac{1}{C}\sum_c \hat{\theta}_c^r\right)^2}{C-1} - v^2 \,.$$

3.3.2 Likelihood Model Approach

An approach that overcomes the above mentioned problems of not obtaining a confidence interval for σ_B^2 and not considering that σ_B^2 has to be estimated (leading to confidence intervals for the overall treatment difference that are too narrow) is the profile likelihood approach (Hardy and Thompson 1996 [18]). It will therefore be utilized for the construction of the confidence intervals.

A normal distribution has to be assumed stringently. According to Whitehead and Whitehead (1991 [53]), the marginal distribution of $\hat{\theta}_c^r$ is normal with mean θ^r and variance $v_c^2 + \sigma_B^2$. After taking partial derivatives of the log likelihood function

$$l(\theta^r, \sigma_B^2) = -\sum_{c=1}^{C} \frac{1}{2}\, ln\left[2\pi\left(v_c^2 + \sigma_B^2\right)\right] - \sum_{c=1}^{C} \frac{(\hat{\theta}_c^r - \theta^r)^2}{2(v_c^2 + \sigma_B^2)}$$

and equating them to 0, the following MLEs are obtained

$$\hat{\theta}^r = \frac{\sum_{c=1}^{C} \frac{\hat{\theta}_c^r}{v_c^2 + \hat{\sigma}_B^2}}{\sum_{c=1}^{C} \frac{1}{v_c^2 + \hat{\sigma}_B^2}} = \frac{\sum_{c=1}^{C} w_c^{\star,r}\, \hat{\theta}_c^r}{\sum_{c=1}^{C} w_c^{\star,r}} \quad \text{(as before)},$$

$$\hat{\sigma}_B^2 = \frac{\sum_{c=1}^{C} \frac{\left(\hat{\theta}_c^r - \hat{\theta}^r\right)^2 - v_c^2}{\left(v_c^2 + \hat{\sigma}_B^2\right)^2}}{\sum_{c=1}^{C} \frac{1}{\left(v_c^2 + \hat{\sigma}_B^2\right)^2}} \,.$$

The two equations have to be solved iteratively beginning with the first equation and an initial value of $\hat{\sigma}_B^2$. It is sometimes argued that the result of the iterative procedure

depends on the starting value chosen. The way the between-center-variance σ_B^2 is estimated with the likelihood model approach differs from the proceeding suggested by DerSimonian.

Likelihood based confidence intervals for θ^r and σ_B^2 can be acquired with the help of profile log likelihoods (Hardy and Thompson 1996 [18]). As mentioned before, profile log likelihoods are the usual log likelihoods for a parameter of interest taking into account that a second parameter is not known and has to be estimated. The profile log likelihood for θ^r is defined as

$$l_1^\star(\theta^r) = l(\theta^r, \hat{\sigma}_B^2(\theta^r)) \,,$$

where $\hat{\sigma}_B^2(\theta^r)$ is the MLE of σ_B^2 with varying θ^r. The profile log likelihood for σ_B^2 is

$$l_2^\star(\sigma_B^2) = l(\hat{\theta}^r(\sigma_B^2), \sigma_B^2) \,,$$

where $\hat{\theta}^r(\sigma_B^2)$ is the MLE of θ^r with varying σ_B^2, respectively. A 95% confidence interval for each of the parameters is then given by all values that satisfy (Hardy and Thompson 1996 [18])

$$\begin{aligned} \text{CI for } \theta^r: \quad & l_1^\star(\theta^r) > l_1^\star(\hat{\theta}^r) - \frac{3.84}{2} \\ \text{CI for } \sigma_B^2: \quad & l_2^\star(\sigma_B^2) > l_2^\star(\hat{\sigma}_B^2) - \frac{3.84}{2} \,, \end{aligned} \tag{3.21}$$

with 3.84 being the 5% quantile of the χ_1^2-distribution.

An alternative way to obtain a confidence interval for σ_B^2 was proposed by Biggerstaff and Tweedie (1997 [4]). The authors calculate an approximate confidence interval for σ_B^2 with the help of an approximate gamma distribution for the test statistic Q. Next to the confidence interval for σ_B^2, Biggerstaff and Tweedie further propose a new estimator for θ in defining a "density form of $\hat{\sigma}_B^2$ as the sum of its discrete and absolutely continuous parts, each weighted by its relative contribution

to the overall distribution of $\hat{\sigma}_B^2$". This results in new weights and hence in different treatment mean estimates. The method incorporates the variability of $\hat{\sigma}_B^2$ in a natural way. The weights proposed by Biggerstaff et al. are approximately lying between the Type III weight of the fixed analysis and the standardized weighting factor of the random approach (3.14). However, this is not to be expected in general according to the authors.

Hartung et al. (2002 [24]) recently investigated the moment estimator $\hat{\sigma}_B^2$ of DerSimonian and Laird in more detail. They found out that in case of 3 participating centers, the moment estimator has a positive bias for $\sigma_B^2 < 0.5$. A small between-center-variance is therefore overestimated. For $\sigma_B^2 \geq 0.5$ the bias is often negative (σ_B^2 is underestimated). If 6 centers are participating in the trial, the moment estimator will commonly overestimate the true between-center-variance. However, this is not always the case. For the trial settings of the simulation study performed in the next chapter, (6 centers, varying sample sizes among the centers, equal within-center-variability), the true between-center-variability is overestimated for $\sigma_B^2 < 10$ but underestimated for $\sigma_B^2 = 10$.

The authors further introduced two new estimators of the between-center-variability. Both are positive and offer the possibility of calculating a confidence interval. Bias and standard deviation of the new estimates are however close to the moment estimate of DerSimonian and Laird. Because the new estimates as well as their confidence intervals are much easier to implement than the iterative REML procedure, the estimates by Hartung seem to be a good choice in case a confidence interval is required. If a confidence interval is not necessary, the DerSimonian moment estimator remains the easiest approach.

3.4 Test of overall treatment effect

For the investigation of the nullhypothesis of no overall treatment effect ($H_0: \quad \theta^r = 0$), the following t-statistic is appropriate. It is designed in the same way as the statistic of a fixed effects model

$$T^r = \frac{\hat{\theta}^r}{\sqrt{\widehat{Var}(\hat{\theta}^r)}} = \frac{\frac{\sum_{c=1}^{C} \hat{w}_c^{\star,r}\, \hat{\theta}_c^r}{\sum_{c=1}^{C} \hat{w}_c^{\star,r}}}{\sqrt{\frac{1}{\sum_{c=1}^{C} \hat{w}_c^{\star,r}}}} = \frac{\sum_{c=1}^{C} \hat{w}_c^{\star,r}\, \hat{\theta}_c^r}{\sqrt{\sum_{c=1}^{C} \hat{w}_c^{\star,r}}}, \quad \text{where} \quad \hat{w}_c^{\star,r} = \frac{1}{\hat{v}_c^2 + \hat{\sigma}_B^2}. \tag{3.22}$$

Hartung and Knapp (2001 [21] [22]) recently proposed to use a different estimator for the variance of $\hat{\theta}^r$. Instead of the variance estimator

$$\widehat{Var}(\hat{\theta}^r) = \frac{1}{\sum_{c=1}^{C} \hat{w}_c^{\star,r}},$$

they suggest to apply

$$\hat{q} = \widehat{Var}(\hat{\theta}^r) = \frac{1}{C-1} \sum_{c=1}^{C} \frac{\hat{w}_c^{\star,r}}{\sum_c \hat{w}_c^{\star,r}} (\hat{\theta}_c^r - \hat{\theta}^r)^2 . \tag{3.23}$$

This estimate can be regarded as a weighted extension of the usual empirical variance of $\hat{\theta}_c^r$ for c=1,...,C. According to the authors, $\hat{q}$ is non-negative, unbiased and invariant with respect to mean value translations.

The Hartung test statistic for a two-sided hypothesis of no overall treatment effect is therefore defined as

$$T^{r,\,Hartung} = \frac{|\hat{\theta}^r|}{\sqrt{\hat{q}}} . \tag{3.24}$$

It will be compared with the $1 - \frac{\alpha}{2}$ quantile of the t-distribution with C-1 degrees of freedom where α represents the level of significance.

In this chapter, mixed effects models have been described so far. Before they will be compared with the fixed effects models of Chapter 2, the connection between multicenter studies and commonly used approaches in meta-analysis will be investigated.

3.5 Meta-Analysis

> *"Meta-analysis is the use of statistical techniques to combine the results of studies addressing the same question into a summary measure" (Villar et al. 2001 [52]).*

The wish for the calculation of a pooled estimator out of different studies is based on the fact that the individual studies are often too small to detect moderate treatment effects. Furthermore, they are often too limited to allow generalization to other patient populations (Whitehead and Whitehead 1991 [53]). As for multicenter studies, a pooled estimate might be misleading in case there are substantial differences between the results of the studies.

The underlying models for meta-analyses are either a mixed effects model or, very rarely, a fixed effects model. The employed models are therefore the same as for multicenter trials. The main difference is that meta-analyses investigate the results of different studies, whereas multicenter trials focus on the different centers.

If a model with fixed effects only is sometimes applied in meta-analyses, the Type III sums of squares where the trials are equally weighted will never be applied (Senn 2000 [49]). Senn states that "no researcher ever uses an approach to analysis in which mega-trials and single center studies are given equal weight".

If both the meta-analysis and the multicenter study are based on a mixed effects model, less heterogeneity will be expected for the multicenter study. Firstly because there is only one global study protocol for all of the centers enrolled in a prospective

multicenter study. For retrospective meta-analyses, a different study protocol exists for each of the trials usually conducted at different times. Secondly, treatment as well as primary outcome variables differ in general for meta-analyses but are the same for multicenter studies. Randomization procedures might be different as well.

The random approach which has just been described was originally designed for meta-analysis. This is the reason why most of the literature concerning random effects is embedded in meta-analyses. In Chapter 3, the formulas and statements for meta-analyses have already been applied at multicenter trials.

The test for heterogeneity (3.20) is sometimes used in meta-analysis to decide on the method (fixed or random approach). If heterogeneity can be concluded from the result of the test, a random approach would be chosen. However, the test usually has low power. A deeper investigation of this topic was performed by Hardy and Thompson (1998 [19]). It revealed that the test for heterogeneity is of limited use particularly when total information (=total weight or inverse variance) is low, or when the amount of information in each trial varies greatly.

After the description of the theoretical bases of fixed and mixed effects models, Chapter 4 is devoted to their comparison. This will be done by giving a short summary of their theoretical peculiarities at the beginning of the chapter as well as by performing simulation studies afterwards.

Chapter 4

Simulation of treatment effects

4.1 Comparison of fixed and random center effects

Regarding the distributional conditions, the fixed effects model and the mixed effects model vary in the following respect: for the fixed effects model homogeneity of treatment effects over all centers is commonly assumed in order to get an unbiased estimate of the overall treatment effect. It is therefore anticipated that each center treatment mean estimator has the same common mean θ but the individual variance v_c^2

$$\hat{\theta}_c \quad \sim \quad N(\theta, v_c^2)$$

assuming normality. In contrast, regarding the mixed effects model, homogeneity of treatment effects is not expected. Under the assumption of normality, the marginal distribution of

$$\hat{\theta}_c \quad \text{is} \quad F_{\hat{\theta}_c} = N(\theta, v_c^2 + \sigma_B^2) \; .$$

The conditional distribution (under δ_c) is

$$\hat{\theta}_c \sim F_{\hat{\theta}_c \mid \delta_c} = N(\theta_c,\, v_c^2)\ .$$

The choice of one of the two methods is not an easy task. A possible proceeding was proposed by Kaufmann (2001 [30]). It is based on the pre-test commonly employed for meta-analyses (Figure 4.1).

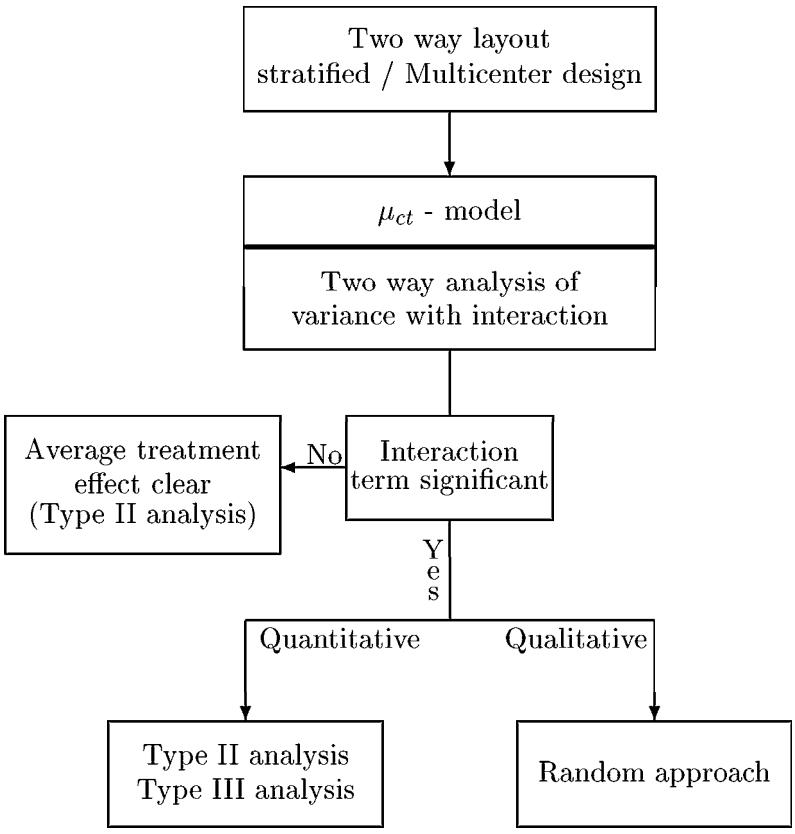

Figure 4.1: Decision on fixed or random approach

When the nullhypothesis of no interaction has to be accepted, the treatment effects of the different centers are assumed to be the same, leading to a well defined overall treatment effect of the magnitude of the equal center treatment effects. As previously

mentioned, the treatment-by-center interaction effect should in this case not be included into the model so that there is no reduction in efficacy (Phillips et al. 2000 [37]). If, however, treatment-by-center interaction is concluded, a fixed approach will be suggested when quantitative interactions are present while a random approach will be proposed if there are qualitative interactions. The choice between the Type II and Type III estimators in case of quantitative interactions remains unsolved.

The employment of pre-testing by Kaufmann is in contrast to the proceeding by Jones et al. (1998 [29]) described in Chapter 2. There, the authors limited the application of a pre-test on a fixed effects model: they suggested to apply Analysis II when interactions cannot be concluded from the pre-test and to use Analysis III if there are interactions. Based on the reasons presented earlier, pre-testing is not regarded as an efficient method and will therefore not be suggested in this writing. Before a better recommendation can be given, some further aspects need to be considered.

A random effects estimator is lying between the Type II and Type III estimator of a fixed effects model (Senn 1998 [48], Jones et al. 1998 [29], Gallo 2000 [14], Hardy et al. 1996 [18]). This is due to the fact that if there is no between-center-variation, which means $\sigma_B^2 = 0$, the random effects estimator will reduce to the Type II overall treatment mean estimator since

$$w_c^r = \frac{\frac{1}{\sigma^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})+0}}{\sum_c \frac{1}{\sigma^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})+0}} = \frac{\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}}{\sum_c \frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}}.$$

On the other hand, if σ_B^2 tends to infinity, the weights of the random approach tend to equality

$$w_c^r = \frac{\frac{1}{\sigma^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})+\infty}}{\sum_c \frac{1}{\sigma^2(\frac{1}{n_{cA}}+\frac{1}{n_{cB}})+\infty}} \quad \longrightarrow \quad \frac{1}{C}.$$

The overall treatment mean estimate of the random effects model will therefore approach the Type III estimator of the fixed effects model if $\sigma_B^2 \to \infty$ (Hardy and

Thompson 1996 [18]). This fact can also be regarded as one of the disadvantages of the DerSimonian and Laird method. If between center variance increases and weights are hence getting closer to equality, the method will assign too much weight to small centers with large variances, and too little weight to large centers with small variances respectively (Biggerstaff et al. 1997 [4]).

Finally, Hartung and Knapp (2001 [22]) propose to use test statistic (3.24) with the new variance estimate $\hat{q}$ (3.23) in general when investigating the hypothesis of an overall treatment effect. The performance of this test with regard to its significance level has proved to be excellent. In contrast, the commonly employed Analyses II and III as well the traditional random approach can be very liberal, especially for large between-center-variabilities. Hartung et al. concluded that the suggested test statistic keeps the prescribed level of significance irrespective of the sample sizes, the number of centers, the presence of homogeneous or heterogeneous variances of the treatment responses within a center as well as irrespective of the magnitude of the unexplained (between-center) variability. Since the test is almost invariant to changes in the true between-study-variance, the authors further deduce that a choice between the fixed effects and random effects approach in advance is not necessary.

However, Hartung and Knapp (2001 [22]) based their recommendation solely on significance level properties of the statistics. Other properties, like the power, have not been investigated yet.

In the next section, all current approaches will be compared within the framework of simulation studies. Power and significance level of the tests based on the two fixed effects models (Type II and Type III analyses) and the two mixed effects models (traditional random approach and random approach by Hartung) will be evaluated. Moreover, data will be assumed to come from a fixed and from a mixed effects model. Power and significance level results will therefore be interpreted separately for both types of data in order to find the more suitable model for each situation.

4.2 Simulation of data from fixed and mixed effects models

Data are therefore simulated from a fixed as well as from a mixed effects model. The simulations are done according to the settings described in Jones et al. (1998 [29]). It is assumed that 6 centers are participating in a trial with the following distribution of patients (Table 4.1)

Table 4.1: Distribution of patients of first simulation study

	Center						
Treatment	1	2	3	4	5	6	$\sum$
t=A	5	5	10	10	15	15	60
t=B	5	5	10	10	15	15	60
$\sum$	10	10	20	20	30	30	120

Concerning the *fixed effects simulation*, the treatment effects θ_c for centers 1 to 5 are assumed to have the value 10, whereas the treatment effect for the sixth center is chosen as +10, +5, 0 and -5. A situation can thus be created in which

Case 1: all centers have equal treatment effects

Case 2: one center has a smaller treatment effect in comparison to the other centers

Case 3: one center has no treatment effect, and

Case 4: one center has a reversed treatment effect.

With regard to the *mixed effects simulation*, the mean treatment effect $\theta = \beta_A - \beta_B$ is set at a value of 10. The treatment-by-center interaction terms $(\alpha\beta)_{ct}$ are simulated from a $N(0, \frac{1}{2}\sigma^2_{\alpha\beta}) \overset{(3.9)}{=} N(0, \frac{1}{4}\sigma^2_B)$ distribution. Center treatment effects

$$\theta_c = \theta + \delta_c = \beta_A - \beta_B + (\alpha\beta)_{cA} - (\alpha\beta)_{cB}$$

result which are $N(10, \sigma_B^2)$ distributed. The between-center-standard-deviation σ_B is chosen as 1, 2, 4, 7, 10 and 20.

For both the fixed and the mixed effects model, the error terms e_{ctk} are simulated from a normal distribution with mean 0 and standard deviation 14. Concerning the mixed effects model, the simulated treatment effects then have the following distributional properties

$$\hat{\theta}_c = \theta_c + \frac{1}{n_{cA}} \sum_{k=1}^{n_{cA}} e_{cAk} - \frac{1}{n_{cB}} \sum_{k=1}^{n_{cB}} e_{cBk} \quad \sim \quad N(10, \sigma_B^2 + 14^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}))\,,$$

where σ_B=1, 2, 4, 7, 10, 20 and n_{ct} as in Table 4.1.

After data have been simulated both from fixed and mixed effects models, the test statistics T^{II}, T^{III}, T^{r}, $T^{r,Hartung}$ for the investigation of an overall treatment effect were computed. Within- and between-center-variability were estimated based on the method of moments. More precisely, the DerSimonian and Laird moment estimator (1986 [7]) of the between-center-variability σ_B^2 (Chapter 3.3.1) has been chosen because it is still widely used. This is probably due to its ease of computation, although the method does not always deliver a precise estimate $\hat{\sigma}_B^2$. Hartung and Knapp (2001 [22]) applied the DerSimonian and Laird moments estimator in their work because the significance level of their proposed test (3.24) is almost invariant with respect to the between-center-variance estimate.

Significance level and power of the four test statistics are presented in subsequent tables. When $|T| > t_{df,1-\frac{\alpha}{2}}$ the two-sided nullhypothesis of equal treatment means ($H_0 : \sum_c w_c \mu_{cA} = \sum_c w_c \mu_{cB}$) was rejected and the alternative of unequal means concluded. With regard to the t-quantile, the significance level α was chosen as 0.05. The degrees of freedom (df) were specified according to the statistic under investigation. Significance level and power were estimated by averaging over 1 million simulation runs.

4.2.1 Results

For the evaluation of the significance level, θ of the mixed effects simulation and θ_c (c=1,...,C) of the fixed effects simulation have been set to zero. Significance level results are presented in Table 4.2.

Table 4.2: Significance level of first simulation study

	Fixed Analyses		Mixed Analyses	
Simulation	Type II	Type III	Mixed	Mixed Hartung
Fixed				
$\theta_c = 0 \, \forall c$	0.05	0.05	0.04	0.05
Mixed				
$\sigma_B = 1$	0.05	0.05	0.04	0.05
$\sigma_B = 2$	0.06	0.05	0.04	0.05
$\sigma_B = 4$	0.08	0.07	0.05	0.05
$\sigma_B = 7$	0.13	0.11	0.07	0.05
$\sigma_B = 10$	0.21	0.17	0.09	0.06
$\sigma_B = 20$	0.45	0.39	0.11	0.05

Concerning data simulated from a fixed effects model, all analyses hold the prescribed level of significance of 5%. When data are assumed to come from a mixed effects model, the fixed effects analyses lead to an inflating level of significance for increases in the variability between the centers. The Type II analysis even has a Type I error of 45%. The two mixed effects analyses perform much better. The approach by Hartung is the only type of analysis which approximately holds the prescribed level also for increases in the between-center-variability (Hartung et al. 2001 [21] [22]). The better performance of the mixed analyses are due to the fact that the estimated variances of the estimated overall treatment effect ($\widehat{Var}(\hat{\theta}^r)$) in

the denominator of the random test statistics T^r, $T^{r,Hartung}$ (3.22)(3.24) are a much better approximation of the true variability than the variances of the estimated overall treatment effect of the fixed effects analyses in the denominator of the test statistics T^{II}, T^{III} (2.32). The latter variances do not include a parameter for the between-center-variability and therefore underestimate the true variability. Larger test statistics are the result.

The above example demonstrates an extreme case. If the number of centers is increased for data based on a mixed effects model the significance levels will decrease so that the difference between the significance levels of the mixed and the fixed approaches reduces. Increasing the number of center e.g. to 20, based on the same overall number of patients and a between-center-variability σ_B=20, results in a level of significance of 28% for the Type II analysis, 17% for the Type III analysis, 7% for the traditional mixed approach and 6% for the Hartung statistic.

The excellent significance level property of the Hartung test is the reason why Hartung generally recommends his approach. However, Hartung has not investigated yet the power performance of his statistic. The power of testing for a treatment difference is presented in Table 4.3. Jones et al. (1998 [29]) already showed power results based on test statistics T^{II}, T^{III} and T^r. The power of the Hartung statistic has been investigated additionally.

Data simulated from a mixed effects model will this time be evaluated first. In accordance with Jones et al., the Type II analysis has the highest power for the whole range of between-center-variability. The power of the Type III analysis is in this example very close to the power of the Type II approach. The power of the Hartung statistic is extremely low in comparison to the power of all other approaches. For the extreme case of σ_B=20, power only reaches a value of 0.25.

Regarding data which comes from a fixed effects model, power is similar to the outcome of the mixed effects simulation. The Type II analysis has once again the highest power for no or small interactions. The Type III analysis now performs

Table 4.3: Power of first simulation study

	Fixed Analyses		Mixed Analyses	
Simulation	Type II	Type III	Mixed	Mixed Hartung
Fixed				
$\theta_6 = 10$	0.97	0.94	0.94	0.86
$\theta_6 = 5$	0.92	0.90	0.86	0.74
$\theta_6 = 0$	0.83	0.83	0.68	0.49
$\theta_6 = -5$	0.68	0.75	0.40	0.22
Mixed				
$\sigma_B = 1$	0.97	0.94	0.94	0.86
$\sigma_B = 2$	0.97	0.94	0.93	0.84
$\sigma_B = 4$	0.96	0.92	0.90	0.79
$\sigma_B = 7$	0.93	0.90	0.82	0.67
$\sigma_B = 10$	0.89	0.86	0.71	0.54
$\sigma_B = 20$	0.78	0.76	0.40	0.25

best for a qualitative interaction although the power of the Type II analysis is only slightly smaller. Power of the Hartung test statistic is still remarkably low. It also highly depends on the magnitude of the interaction.

It is of course of primary importance that a test keeps the prescribed level of significance. When the data are based on a fixed effects model, all analyses hold the level of significance and the fixed effects approaches - especially the Type II analysis - are often much more powerful.

For data based on a mixed effects model, the Hartung approach should be used when the variability between the centers is large because it is the approach with the best significance level properties. A very low power must however be expected. All other approaches do have an inacceptable level of significance. If the between-

center-variability is regarded as being moderate and if an increase in the significance level of 1-2% is not of great concern, the random approach may be an appropriate solution. In comparison with the Hartung statistic, the usual random approach obtains for $\sigma_B = 4$ a power which is 11% higher based on the same significance level. Another example is a power which is 15% higher based on a significance level which is only 2% larger ($\sigma_B = 7$). When the variability between the centers is small or zero, the Type II analysis should again be favored. Hence, the suggestion given by Hartung et al. (2001 [21])) to apply their statistic generally is not recommended and may only be supported from a very conservative point of view.

With regard to the significance level of data which come from a mixed effects model, a point at issue is the variability between the centers. Some statisticians may argue that under the nullhypothesis - that is in case there is no treatment effect - the variability between the centers can only be considered as small or zero. The fixed effects analyses then approximately hold the 5% level of significance. Overall, the fixed effects analyses would generally be preferred over the mixed effects analyses because they achieve a higher power based on the same level of significance. For our considerations it was assumed that the variability between the centers as well as the variability within the centers should be the same under the nullhypothesis as well as under the alternative.

Since it is often difficult to decide if data come from a fixed or a mixed effects model it is proposed to base the choice between the different types of analyses on the magnitude of the between-center-variability. The between-center-variance is commonly not known and has to be estimated. The method of moments by DerSimonian and Laird described in Chapter 3.3.1 is the preferred approach in practice although it does not always deliver precise estimates. An imprecise between-center-variance estimate may not only affect the choice between the types of analyses but may also influence the outcome of the mixed effects analyses because $\hat{\sigma}_B^2$ is part of the random weighting factor. The influence of an imprecise variance estimator on the outcome of the mixed effects analyses will be investigated in the next section

within the framework of sensitivity analyses.

Additionally, the fixed effects model will be further examined in Section 4.3. The Type II and Type III weighting factors will be connected via certain functions in order to investigate the influence of intermediate weights on the outcome of the fixed effects analyses.

4.3 Weight engendering functions

In this section the Type II and Type III weighting factors of the fixed effects model are connected by an exponential function (Gallo 2000 [14]) in order to evaluate the effect of other weighting factors lying between the two "extremes". This can be regarded as a special kind of sensitivity analysis. Hence, weight engendering functions $w_c^\star(u)$, $0 \leq u \leq 1$ connecting

$$w_c^\star(0) = \frac{n_{cA} * n_{cB}}{n_{cA} + n_{cB}} \quad \text{and} \quad w_c^\star(1) = 1 \tag{4.1}$$

exponentially are considered. Data is therefore examined based on a parameter u. Values of u outside this range are of no practical importance.

In a next step a sensitivity analysis of the mixed effects model is performed. The moment estimator of the between-center-variability σ_B^2 is varied between zero and infinity in order to evaluate the influence of an imprecise variance estimate on the result of both mixed effects analyses. The outcome of the mixed approaches is therefore defined as a function of the between-center-variability. The function will be rewritten so that it depends on a scaled between-center-variance and therefore on a parameter u which is also varied between 0 and 1.

4.3.1 Fixed center effect

Exponential function

When the Type II and Type III weighting factors of (4.1) are connected exponentially, the resulting weighting function looks as follows (Gallo 2000 [14]),

$$w_c^{\star,exp}(u) = \left(\frac{n_{cA} * n_{cB}}{n_{cA} + n_{cB}}\right)^{(1-u)} = \frac{1}{\left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right)^{1-u}}, \qquad \text{with} \quad 0 \leq u \leq 1$$

$$\Rightarrow w_c^{exp}(u) = \frac{w_c^{\star,exp}}{\sum_{c=1}^{C} w_c^{\star,exp}} = \frac{\left(\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}\right)^{(1-u)}}{\sum_{c=1}^{C}\left(\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}\right)^{(1-u)}} . \quad (4.2)$$

The estimation function of the difference in treatment means over all participating centers then is

$$\begin{aligned}\hat{\theta}^{exp}(u) &= \sum_{c=1}^{C} w_c^{exp}(u)\,\hat{\theta}_c \\ &= \frac{1}{\sum_{c=1}^{C}\left(\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}\right)^{(1-u)}} \sum_{c=1}^{C}\left(\frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}\right)^{(1-u)} * (\overline{Y}_{cA.} - \overline{Y}_{cB.}) . \quad (4.3)\end{aligned}$$

Expected value and variance are calculated according to (2.17) and (2.18) with the weighting factors now being weighting functions depending on the parameter u. Determination of the test statistic function $T^{exp}(u)$ is based on (2.19) where the MSE (2.20) is employed as an unbiased estimate of the unknown variance σ^2.

Linear Function

A second choice would be the linear connection of the non-standardized forms of the two extreme weighting factors (4.1)

$$w_c^{\star,lin}(u) = \left(1 - \frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}\right) u + \frac{n_{cA}*n_{cB}}{n_{cA}+n_{cB}}, \qquad \text{with} \quad 0 \le u \le 1 . \quad (4.4)$$

Values of u outside the range [0,1] would again be of no importance.

Connecting weight II and weight III with a linear instead of an exponential function yields of course the same weighting factors, but for different values of u. For the rough evaluation of the robustness of weights II and III and hence of the whole range of intermediate weights, both functions are regarded as being equally appropriate. The exponential connection has been chosen.

4.3.2 Random center effect

The weighting factor for the calculation of the overall treatment mean estimate in a mixed effects model was presented in Chapter 3 (3.14) as

$$\begin{aligned} w_c^{\star,r} &= \frac{1}{\text{within + between center variance}} = \frac{1}{\hat{v}_c^2 + \hat{\sigma}_B^2} \\ &= \frac{1}{\hat{\sigma}^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}) + \hat{\sigma}_B^2} \approx \frac{1}{\widehat{Var}(\hat{\theta}_c^r)} \,. \end{aligned}$$

In the latter formula it was already assumed that both the within- and the between-center-variability are not known. They are estimated according to the method of moments. Since the DerSimonian and Laird moment estimator of the between-center-variability (Section 3.3.1) may be imprecise, the influence of a varying moment estimator on the overall decision of accepting or rejecting the nullhypothesis of no overall treatment effect will be investigated. For this reason, the following weighting function depending on the between-center-variability σ_B^2 is employed (Kaufmann 2001 [30])

$$w_c^{\star,r}(\sigma_B^2) = \frac{1}{\hat{\sigma}^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}) + \sigma_B^2} \qquad \text{for } \sigma_B^2 \in [0, \infty] \,, \tag{4.5}$$

with special interest in σ_B^2 taking on values in the region around its moment estimator $\hat{\sigma}_B^2$. The estimator function of the overall treatment effect is then defined as

$$\hat{\theta}^r(\sigma_B^2) = \sum_{c=1}^{C} w_c^r(\sigma_B^2)\ \hat{\theta}_c = \sum_{c=1}^{C} \frac{\frac{1}{\hat{\sigma}^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}) + \sigma_B^2}}{\sum_c \frac{1}{\hat{\sigma}^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}) + \sigma_B^2}}\ \hat{\theta}_c \,.$$

According to Hardy and Thompson (1996 [18]), "a plot of $\hat{\theta}^r$ against σ_B^2 may be used as a form of sensitivity analysis in order to investigate the robustness of $\hat{\theta}^r$ to changes in the value of σ_B^2". The authors actually plot $\hat{\theta}^r$ against the scaled

between-center-variance $u = \frac{\sigma_B^2}{\hat{\sigma}_B^2 + \sigma_B^2}$. A value of

$$\begin{aligned} \sigma_B^2 = 0 \quad &\text{therefore yields} \quad u = 0\,, \\ \sigma_B^2 = \hat{\sigma}_B^2 \quad &\text{yields} \quad u = \frac{1}{2}\,, \\ \text{whereas for } \sigma_B^2 \to \infty\,, \quad &u \to 1\,. \end{aligned}$$

The weighting function of the mixed effects model *depending on the variable u* is obtained in substituting σ_B^2 in formula (4.5) by $\frac{u}{1-u}\,\hat{\sigma}_B^2$

$$w_c^r(u) \;=\; \frac{w_c^{\star,r}(u)}{\sum\limits_c w_c^{\star,r}(u)} = \frac{\frac{1}{\hat{\sigma}^2\left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) + \frac{u}{1-u}\,\hat{\sigma}_B^2}}{\sum\limits_c \frac{1}{\hat{\sigma}^2\left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) + \frac{u}{1-u}\,\hat{\sigma}_B^2}} \qquad 0 \le u < 1\,. \tag{4.6}$$

At u=1, the weighting function is not defined.

Both weight engendering functions (the fixed function of Section 4.3.1 and the random function of Section 4.3.2) are hence evaluated for values of u between 0 and 1. The two functions coincide in the Type II weighting factor at u=0. The exponential weighting function reduces to the Type III weighting factor at u=1, while the random weighting function tends to the Type III weight as u tends to 1.

With regard to the sensitivity analysis of the mixed effects model, the weighting function (4.6) will be the basis for the t-statistic functions $T^r(u)$ and $T^{r,Hartung}(u)$, calculated according to (3.22) and (3.24) respectively. The only difference is that the statistics now depend on the parameter u.

For subsequent simulation study, the settings used in Jones et al. (1998 [29]) are again applied to allow for comparisons with the former investigations. Hence, six centers are assumed to participate in the simulated example with a distribution of patients as presented in Table 4.1. This time, data were assumed to come from a mixed effects model only. The assumptions for the mixed effects simulation ($\theta = 10$, $\sigma_B = 10$, $\sigma = 14$) are the same as described at the beginning of Chapter 4.2. Power and significance level of a 2-sided test for an overall treatment effect are based on 500,000 simulation runs.

4.3.3 Results of sensitivity analyses

In Figure 4.2, the significance level depending on the parameter u is presented for the exponential weight engendering function of a fixed effects model as well as for the weighting functions of both random approaches.

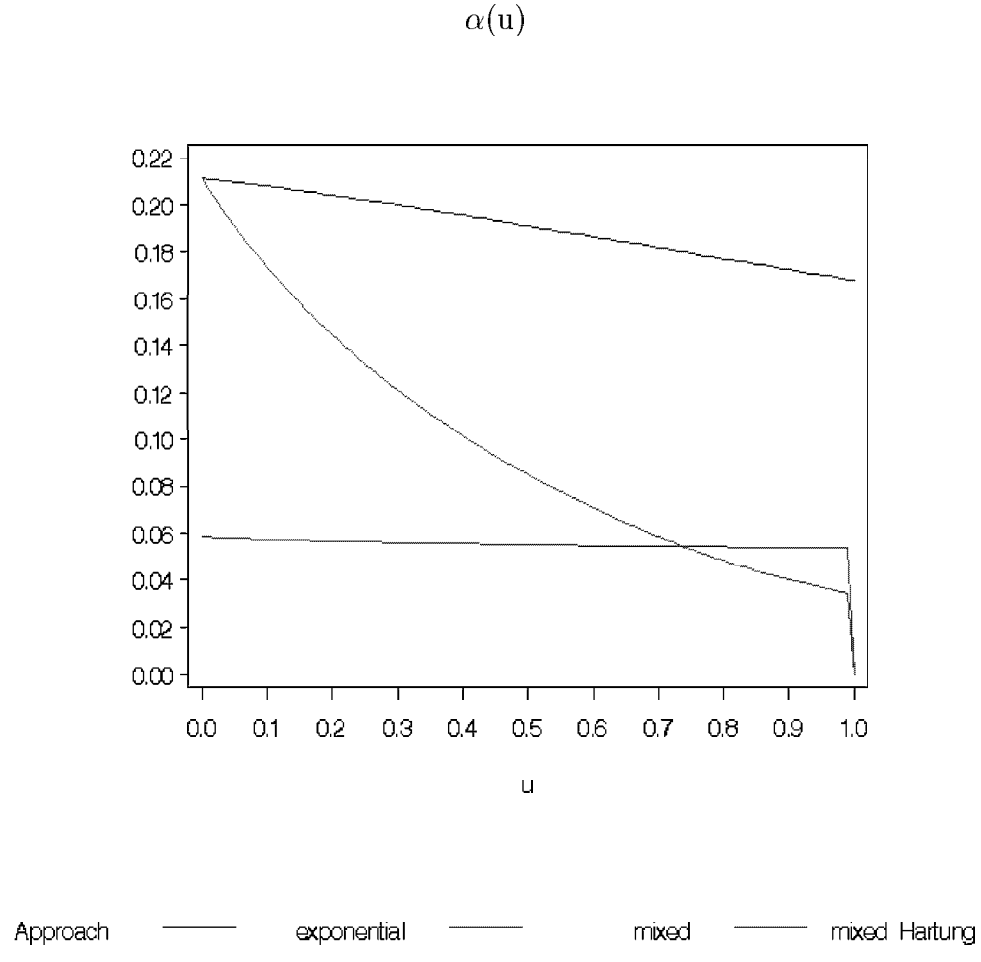

Figure 4.2: Significance level functions of first simulation study

With regard to the fixed effects model, the significance level at u=0 corresponds to the significance level of the Type II analysis. At u=1, the significance level of the Type III analysis is obtained. As can be seen from the figure, the significance level of the fixed effects model is strictly monotonic decreasing from 0.21 at u=0 to 0.17

at u=1. Hence, intermediate weighting factors do not lead to any surprising results in this example.

The traditional mixed effects analysis coincides with the Type II analysis at u=0 which corresponds to a between-center-variability of 0. With increasing u (increasing σ_B^2) the significance level drops steeply from 0.21 to a significance level of around 0.03, before it finally tends to 0 for a variance that approaches infinity. The significance level of the traditional mixed approach is therefore highly affected by an imprecise estimate of the between-center-variance in the region around the DerSimonian and Laird estimator (u=0.5). In contrast and as already stated by Hartung et al. (2001 [21] [22]), the significance level of their statistic is almost invariant to changes in the moment estimator of the between-center-variability. With the performed sensitivity analysis these findings could be well confirmed. In fact, the Hartung statistic approximately holds the defined significance level of 5% also for extreme choices of the between-center-variability.

In Figure 4.3 the power depending on the parameter u is presented. In accordance with Table 4.3 ($\sigma_B = 10$), the power of the fixed effects model is 0.89 at u=0 and 0.86 at u=1. The course of the function is first slightly increasing and then decreasing. All weighting factors between weight II and weight III therefore lead to high and similar power values.

The power of the usual random approach corresponds to the power of the Type II analysis of 0.89 at u=0 (σ_B^2=0) and decreases monotonously as the between-center-variability increases. As u tends to 1, power steeply tends to zero. This is due to the fact that in the example under investigation, a value of u that tends to 1 yields a moment estimator $\hat{\sigma}_B^2$ which is multiplied by infinity, since in formula (4.6) $\frac{u}{1-u} \to \infty$ for $u \to 1$. In contrast, a value of u=0.99 does not yet result in an extremely high variance because the moment estimator is only multiplied by $\frac{u}{1-u} = 99$. This is the reason for the sharp bend of the power (and the significance level) functions of the mixed effects models between u=0.99 and a value of u $\to$ 1. At $u = \frac{1}{2}$ (corresponding to the DerSimonian and Laird moment estimator of

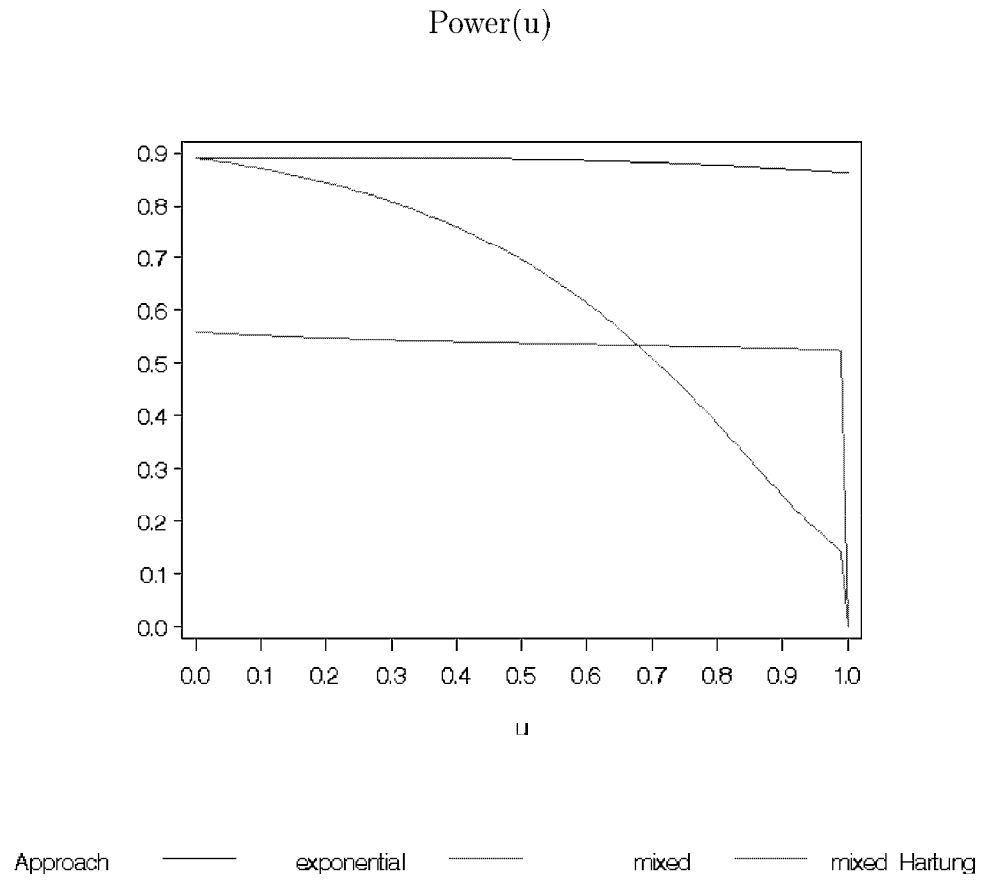

Figure 4.3: Power functions of first simulation study

the between-center-variance), a power of 0.71 is obtained for the traditional random approach as already demonstrated in Table 4.3. In the area around the moment estimator, the power of the traditional random approach changes remarkably. The moment estimator therefore seems to have a great influence on the significance level and the power of the statistical test of the traditional random approach.

In contrast, the Hartung statistic leads in this example to an approximately constant power of around 0.54. The course of the power function is slightly decreasing for $0 \leq u \leq 0.99$ ($0 \leq \sigma_B^2 \leq 5723$) before it plunges down to zero for $u \to 1$. Overall, the Hartung statistic yields a lower power than the usual random approach in the region around the moment estimate $\hat{\sigma}_B^2$ ($u = \frac{1}{2}$). However, it stays almost constant for changes in $\hat{\sigma}_B^2$.

In order to investigate in more detail if not only the significance level but also the power of the Hartung statistic is generally invariant with regard to the between-center-variance, further simulation studies were performed. As an example, the number of patients at the different centers was varied in the former study resulting in greatly unequal center sizes. Hence instead of the distribution of patients presented in Table 4.1, the sample sizes of Table 4.4 were considered.

Table 4.4: Distribution of patients of second simulation study

	Center						
Treatment	1	2	3	4	5	6	$\sum$
t=A	1	7	7	10	12	23	60
t=B	3	6	8	8	13	22	60
$\sum$	4	13	15	18	25	45	120

There were again 60 patients in each of the two treatment groups. The distribution of the 120 patients across the six centers had this time been simulated. In order to simulate a random distribution of patients across centers it is generally assumed that the recruiting of new patients to a center c follows a Poisson process with parameter λ_c, c = 1,...,C (Lin 1999 [34]). The probability that a newly recruited patient belongs to center c is $\Pi_c = \frac{\lambda_c}{\sum_{c=1}^{C} \lambda_c} \; \forall c = 1, ..., C$. Since the λ_c are unknown and may vary between the centers and from study to study, they are assumed to be independent, identically distributed and follow the exponential distribution with parameter 1. The probability vector "$(\Pi_1, ..., \Pi_C)$ then follows a Dirichlet distribution with all parameters equal to 1" (Lin 1999 [34]). In order to obtain the distribution of a sample of N patients among C centers, the observed probabilities $(\Pi_1, ..., \Pi_C)$ are used as parameters of a multinomial distribution.

Response data were again assumed to come from a mixed effects model with $\theta = 10$ and $\sigma = 14$. Between-center-standard-deviation σ_B was anticipated to have the value 7. Significance level and power functions are presented in Figure 4.4 and Figure 4.5.

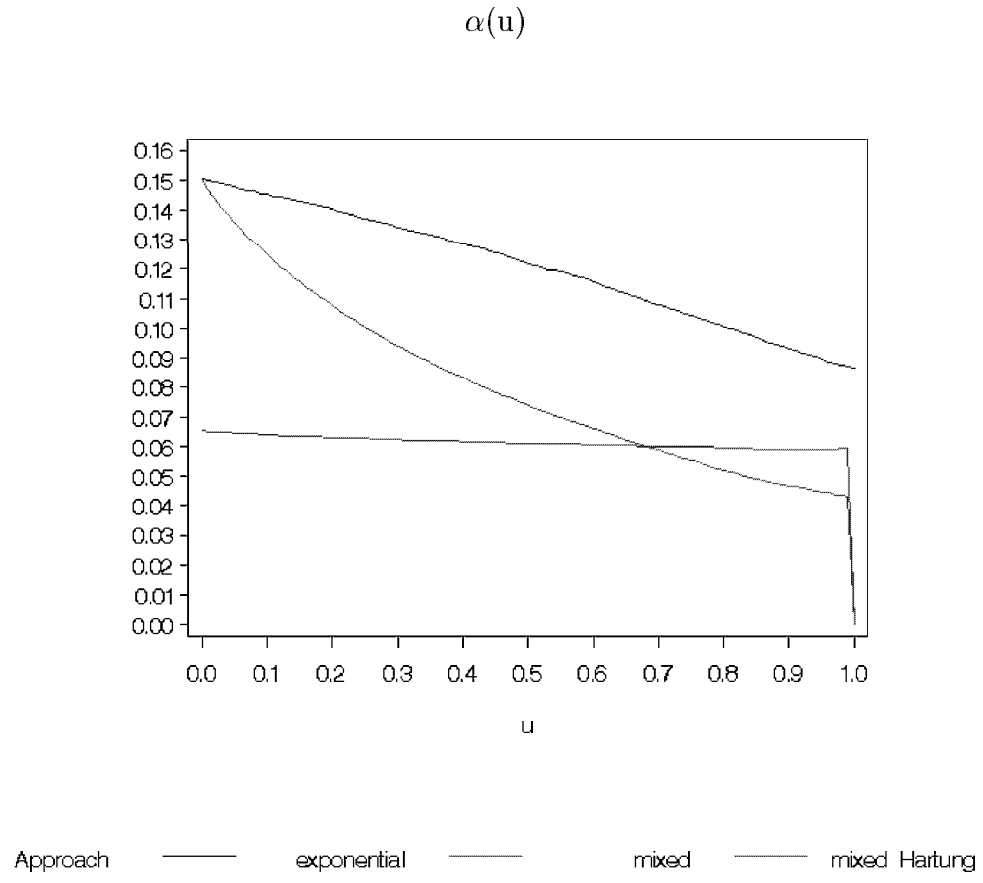

Figure 4.4: Significance level functions of second simulation study

The significance level of the Type II analysis at u=0 is again worst. It even increased from 13% (Table 4.2) to 15% due to the center sizes which differ much more than in the example of Chapter 4.2. For the same reason the significance level of the Type III analysis at u=1 drops from 11% (Table 4.2) to 9%. The significance level of the Hartung statistic is again almost invariant to changes in the between-center-variability.

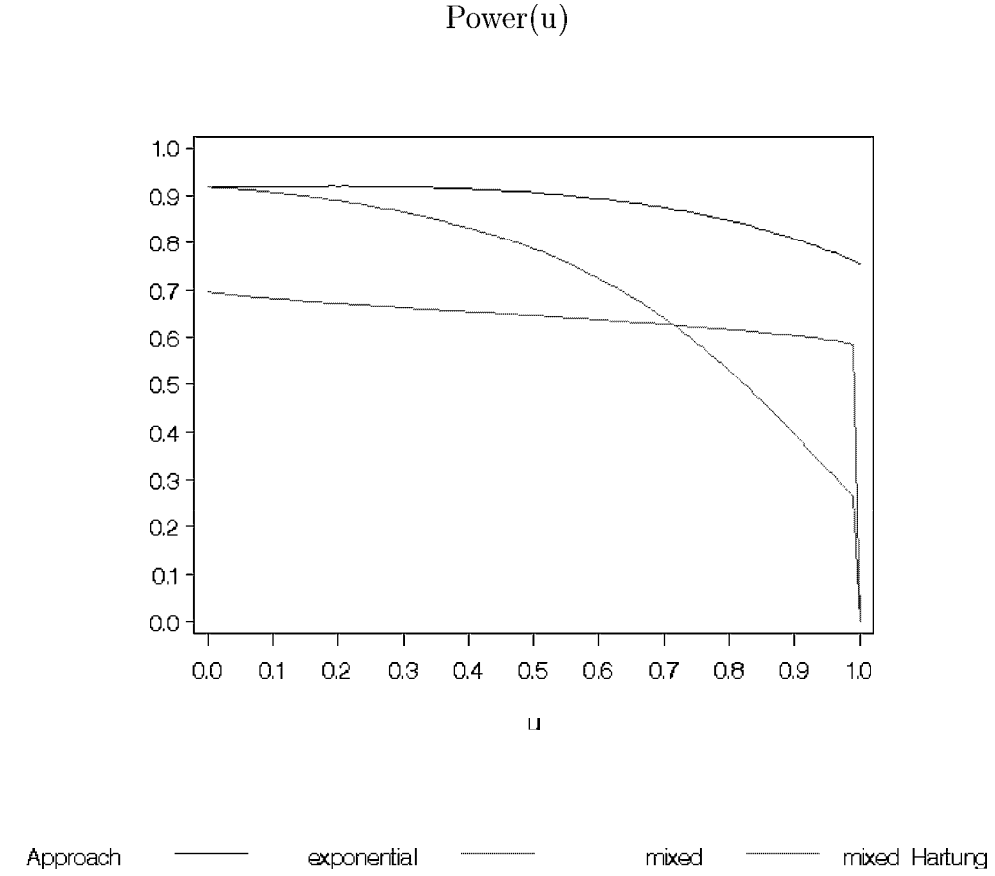

Figure 4.5: Power functions of second simulation study

The greatly varying sample sizes are also responsible for the decrease in power of the Type III analysis at u=1. Its power of 76% is this time lower than the power of the traditional random approach at u=0.5 (79%). The slope of the power function of the Hartung approach is more ascending than in the previous example (Figure 4.3). It decreases from 70% to 58% as u is varied from 0 to 0.99 ($\sigma_B^2 = 0$ to 3403). The Hartung statistic is therefore less invariant to changes in σ_B^2 than in the first simulation.

The robustness of the power of the Hartung statistic will also be investigated for data which are assumed to come from a fixed effects model. The example is based on the Transpulmin Study which will be described in detail in Chapter 5. The simulation demonstrates that the power of the Hartung approach can sometimes be

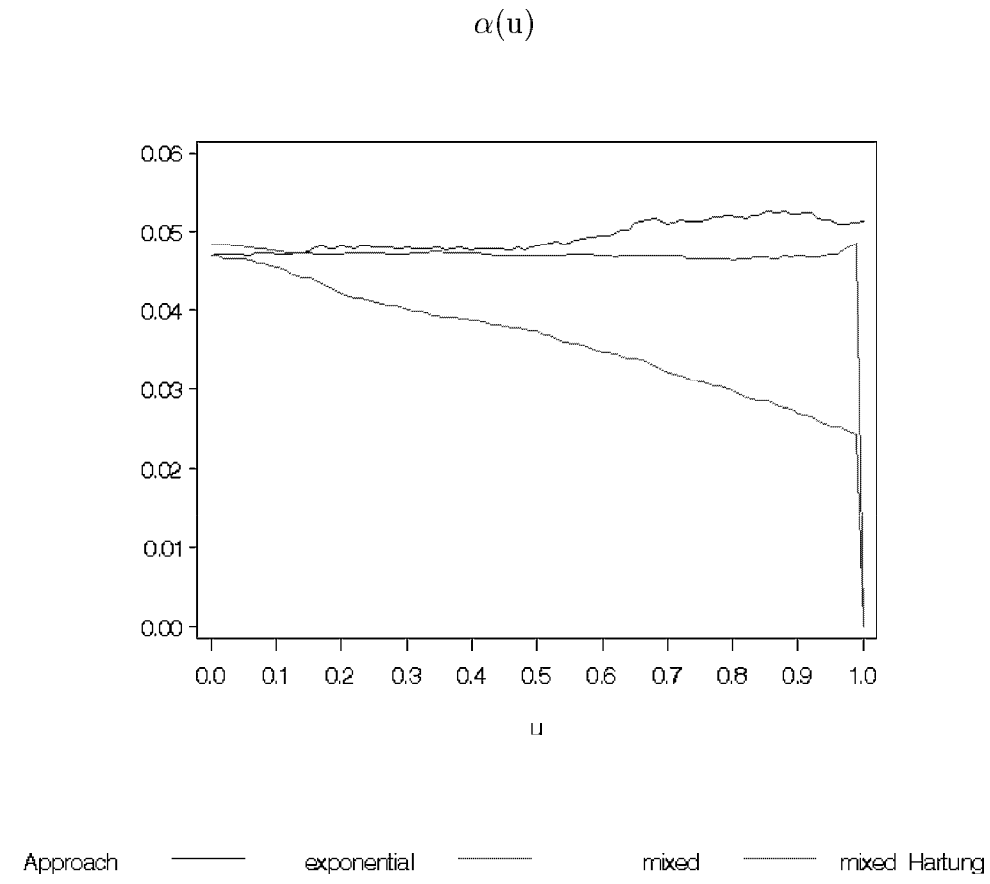

Figure 4.6: Significance level functions of Transpulmin Study

greatly affected by an imprecise estimate of the between-center-variability. Thirteen centers participated in that trial with a total number of patients of 256 (Table 5.1). It was assumed that normally distributed responses come from a fixed effects model with the following center treatment effects: +2.0, 0, -2.0, 0, +2.5, +0.5, -2.0, +4.0, +1.0, +3.5, +0.5, 0, +0.5. Within-center-variability σ^2 was chosen to have the value 9.

Since data is based on a fixed effects model, all types of analyses hold the level of significance of 5% (Figure 4.6). The traditional mixed effects model is as usual affected by changes in the estimated variability between the centers. Concerning power, the Hartung approach now varies remarkably for changes in u which is most likely due to the greatly varying center sizes (Figure 4.7). The power variation

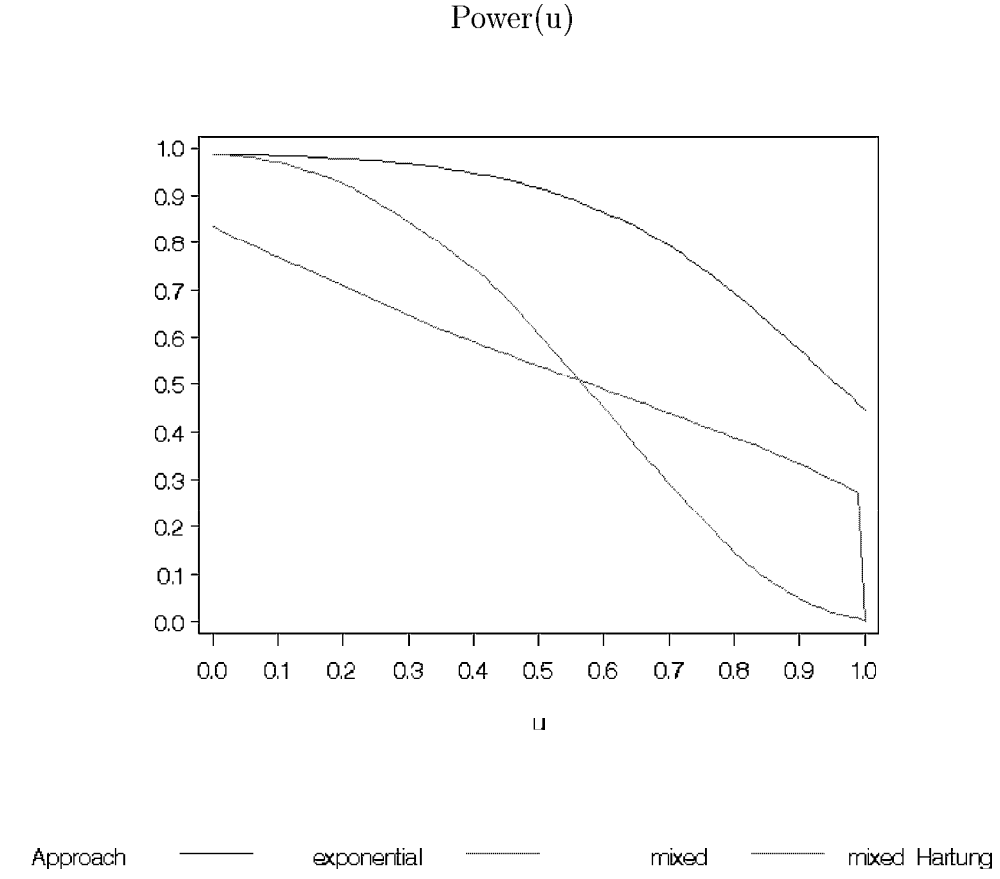

Figure 4.7: Power functions of Transpulmin Study

is however less extreme than the power variation of the traditional mixed effects model. The reason for the smaller variation is the variance estimator of the overall treatment effect (3.24) in the denominator of the Hartung statistic which is less affected by changes in the moment estimator of the between-center-variance than the variance estimate in the denominator of the traditional random approach (3.22). The Transpulmin Study also demonstrates the tremendous loss in power for the Type III analysis (at u=1) when the sample sizes vary greatly between the centers. The power of the Type III analysis (45%) is worst and therefore even lower than the power of the Hartung statistic at u=0.5.

Summarizing the findings of Section 4.3, the exponential connection of the Type II and Type III weighting factors of a fixed effects model led for the examples under

investigation (based on data from a fixed and a mixed effects model) to power and significance level functions which did not always increase or decrease monotonously. However, only small rises (or declines) in the functions could be detected. The connection of the "extreme" weights therefore seems to be a good and additional method of exploring not only the outcome of the Type II and III analyses but also their "robustness" although the connection is not considered as being of primary importance. In cases of greatly varying sample sizes between the centers or of a zero estimate of the between-center-variance, particular store should of course be set on the result of the Type II analysis due to its higher power.

Moreover it could be demonstrated that for data which come from a fixed or from a mixed effects model the traditional random approach is highly affected by the estimate of the between-center-variability. Power and significance level vary widely around $u = \frac{1}{2}$. In contrast, the Hartung statistic stays almost constant with respect to the level of significance. Concerning power, the Hartung statistic proved to depend on the variance between the centers although its power varies less than the power of the traditional random approach.

The suggestion of using the traditional random approach as an alternative to the random approach by Hartung in case of a moderate estimate of the between-center-variability cannot be supported furthermore. A random approach which is highly affected by an imprecise moment estimate might result in an inflated level of significance. The approach by Hartung is the better solution.

Since it is often difficult to decide if data come from a fixed or a mixed effects model it will be recommended to start the evaluation of data from a multicenter trial either with the connection of the Type II and III analyses or their exclusive calculation. Thereby, the sensitivity or conformity of both analyses is investigated. A much higher power for the conclusion of a treatment effect is expected for at least one of the fixed effects analyses in comparison to the mixed model by Hartung. One has to be aware of the fact that the significance level might be inflated. The random approach by Hartung is suggested next. It is expected to hold the level of

significance at the expense of a much lower power which may additionally depend on the estimate of the between-center-variance. In cases where the random and fixed approaches all lead to the rejection (or the acceptance) of the nullhypothesis of no treatment effect, the overall conclusion is clear. Results of the Hartung approach should be presented due to their better significance level properties.

If the conclusions differ, more emphasis should be assigned to the fixed effects analyses if the moment estimator of the between-center-variance is low or even zero. A much higher power based on the same level of significance is achieved in comparison to the random approach by Hartung.

For a moderate to large moment estimator of σ_B, the Hartung approach should obtain more weight in order to keep the prescribed level of significance. If the choice between the two approaches is not clear due to the possible imprecision of the moment estimator, either the complex calculation of the maximum likelihood estimator of the between-center-variability, its confidence interval based on profile likelihoods or the new variance estimates by Hartung et al. (2002 [24]) might be of help.

A completely new weighting factor will be introduced in the next chapter. It will also be applied to the Transpulmin Study. The idea of the new weighting factor is based on the definition of a standard center in advance of the study. The participating centers will be compared with this "standard". Their deviation will be quantified by means of a divergence measure.

Chapter 5

The standard center

5.1 Divergence measures

5.1.1 Kullback-Leibler Number

The Kullback-Leibler number, also known as relative entropy, has its roots in information theory. It can be interpreted as a generalization of the Shannon entropy (Shannon 1948 [50]). Furthermore it belongs to the class of non-parametric information measures and is one of the most popular distance measures in statistics. The Kullback-Leibler number quantifies the "distance" between two distribution functions P and Q which are represented by their density functions p and q (Kullback et al. [33]).

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and let $\mathcal{E} \subset \mathbb{R}$ be a countable set of real numbers so that the random variable X: $\Omega \longrightarrow \mathcal{E}$ is discrete. The Kullback-Leibler number for two discrete densities p and q is defined as

$$D^{KL}(p \,||\, q) = \sum_{i \in \mathcal{E}} p(i)\, ln \frac{p(i)}{q(i)}\,, \tag{5.1}$$

where p(i)=P(X=i) $\forall\, i \in \mathcal{E}$ and q(i) respectively. Kullback et al. (1951 [33]) proved that $D^{KL}(p \,||\, q) \geq 0$ applying Jensen's inequality and the convexity of $t\,ln\,t$. The distance will be zero if and only if p=q.

Strictly speaking, the Kullback-Leibler number does not determine a mathematical distance. In general, the symmetry property ($D^{KL}(p \,||\, q) = D^{KL}(q \,||\, p)$) as well as the triangle inequality ($D^{KL}(p \,||\, r) \leq D^{KL}(p \,||\, q) + D^{KL}(q \,||\, r)$) are not met (Kullback et al. 1951 [33], Kullback 1959 [32]). Since the distance is not symmetric, it is better to interpret $D^{KL}(p \,||\, q)$ as the divergence of p from q. This is the reason why the Kullback-Leibler number is often called directed divergence (discrimination information, entropy distance, cross-entropy).

In spite of the missing symmetry and triangle inequality property, the Kullback-Leibler number is very popular and widely used due to its simplicity and broad applicability (Jensen 1993 [28]).

5.1.2 Matusita Distance

Another common distance measure as well as a familiar tool in decision theory is the Matusita distance (Matusita 1955 [35]). It is defined in its discrete version as

$$D^{M}(p \,||\, q) = \sqrt{\sum_{i \in \mathcal{E}} \left(\sqrt{p(i)} - \sqrt{q(i)}\right)^2} \tag{5.2}$$

and takes on values in the interval $[0, \sqrt{2}]$. In contrast to the Kullback-Leibler number, the Matusita distance meets all distance postulates

- $D(p \,||\, p) = 0$
- $D(p \,||\, q) \geq 0$
- $D(p \,||\, q) = D(q \,||\, p)$

- $D(p \,||\, q) \leq D(p \,||\, r) + D(r \,||\, q)$,

where p, q and r are density functions.

Concerning the Matusita distance, the relation to the Shannon entropy and hence to the basis of information theory is not as close as for the Kullback-Leibler number. Because of its excellent distance properties, the Matusita distance is important as a real distance measure as well as in decision theory. The opportunities of its use are usually rare. However, for the topic under investigation, the applicability and calculation of the Matusita distance does not rank behind the Kullback-Leibler number.

A further advantage of the Matusita distance is its good approximation of the Rao distance (Jensen 1993 [28]). The Rao distance is a very complex measure that changes its shape depending on the underlying distribution. The idea is to regard a n-parametric distribution family as a generalization of a n-dimensional set of real numbers. The Rao distance of two probability distributions then measures the minimum arc length of all curves that connect the distributions (points) (Jensen 1993 [28]). This distance measure has excellent theoretical properties e.g. invariance. It can be regarded as a generalization of the Euclidean distance. However, in order to employ the Rao distance, certain regulatory conditions need to be fulfilled. The class of distributions which basically comes into question is the exponential family. Jensen (1993 [28]) mentioned as a further constraint of the Rao distance that in some situations it is either not possible to calculate it because of mathematical difficulties, or the calculation is consuming a lot of time. Such restrictions do not exist for the Matusita distance and the Kullback-Leibler number. Therefore, the Matusita distance is the preferred distance measure for the new weighting factor which will be introduced shortly.

5.2 On wisdom before the trial

According to DerSimonian et al. (1986 [7]) "one problem in combining studies in meta-analysis is the assignment of weights that reflect the relative *value* of the information provided in a study". For multicenter trials, weights that might better (than the ones up to this point) reflect the information provided by a center will be introduced in this section.

The new approach will be based on a fixed effects model. Heterogeneity between the centers will however still be considered. It will be assumed that the center treatment effects mainly differ due to the varying samples of patients recruited at the different hospitals. For example, some centers may have a very heterogeneous sample of patients while the patients of other centers are more homogeneous. One center may only be able to recruit old patients whereas a second center mainly treats young patients. The different samples of patients will hence be regarded as the main reason for heterogeneity of treatment effects.

Each center should therefore possess a certain and comparable patient sample. For this reason a "standard sample of patients" which belongs to an imaginary "standard center" will be defined in advance of the trial. Their definition will be based on the patient characteristics which are regarded as being of primary importance. The information of a participating center will be regarded as high if it recruits patients with the desired characteristics. Or in other words, if the patients of the participating center correspond to the patients of the "standard center".

The deviation of each patient sample from the "standard sample of patients" will be quantified by the Matusita distance as will be described in detail in the following section. The observed deviation will be included into a new weighting factor so that centers with the desired sample of patients are "upweighted" whereas deviant centers will be "downweighted". Results from differing centers can thus be combined easily. The new weighting factor will now be described at length.

5.2.1 The new weighting factor

It is assumed that I patient characteristics have been chosen to be of primary importance for the description of the desired sample of patients. For our considerations, the characteristics are expected to be of qualitative nature. Concerning the standard center, the actual proportions of study patients with the specific factor levels for each characteristic have to be determined. This might be done by means of population data published by the Federal Statistical Office. The standard center is therefore assumed to have a sample of patients with a particular distribution concerning each of the characteristics. The deviation of the center distributions from the predefined distributions of the standard center will then be quantified based on the Matusita distance.

If a characteristic is continuous, a classification into few classes will be suggested. The classes should roughly reflect the actual patient distribution. The amount of classes should not be large. Otherwise, it will become almost impossible for a center to achieve the targeted distribution, especially if the sample size of a center is small.

It is assumed that the ith characteristic of the standard center ($i = 1, ..., I$) has K_i factor levels (=classes). The probabilities of a random variable X_i falling in each of the classes $k = 1$ to K_i will be defined as $q_{X_i}(k)$. With regard to the participating centers, the probability of the random variable X_{ci} of the cth center ($c = 1, ..., C$) falling into the classes $k = 1$ to K_i will be defined as $p_{X_{ci}}(k)$, respectively.

For each of the I characteristics, the deviation of the center distribution $p_{X_{ci}}$ from the distribution of the standard hospital q_{X_i} will be calculated based on the Matusita distance measure (5.2)

$$d_{ci}^{M}(p_{X_{ci}} \,||\, q_{X_i}) \;=\; \sqrt{\sum_{l=1}^{K_i} \Big[\sqrt{p_{X_{ci}}(l)} - \sqrt{q_{X_i}(l)} \Big]^2} \;, \qquad i = 1\,, \ldots, I\,. \tag{5.3}$$

As previously mentioned, the distance measure takes on values in the interval $[0,\sqrt{2}]$. If the distribution of the ith characteristic of the cth center is the same as the

corresponding distribution of the standard clinic, the measure will yield a distance of 0. In all other cases it is positive. The overall distance of a center from the standard center is obtained as the sum of the I distributional distances

$$d_c = \sum_{i=1}^{I} d_{ci}(p_{X_{ci}} \,||\, q_{X_i}) \qquad\qquad c = 1, \ldots, C\,.$$

This distance takes on values in the interval $[0, I\sqrt{2}]$. There are several possibilities of incorporating the center deviation d_c into a weighting factor. First of all, a weighting factor which is only based on the observed deviation $w_c^\star = \frac{1}{1+d_c}$ could be employed. It will not be suggested because it has been shown in former chapters that weighting factor II and not weighting factor III will perform best if there is no sign of heterogeneity and therefore if $d_c = 0$. Two further possibilities which stand to reason will now be presented.

5.2.2 Possibility one

The distance of center c can be multiplied with the pooled within-center-variability in the following way

$$\begin{aligned} w_c^{\star,V_1} &= \frac{1}{\text{within-center-variability * (1+ center deviation)}} \\ &= \frac{1}{v_c^2 * (1 + d_{c1} + d_{c2} + \ldots + d_{cI})} \\ &= \frac{1}{\sigma^2 \left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) * (1 + d_c)} \qquad (5.4) \end{aligned}$$

$$\Longrightarrow w_c^{V_1} = \frac{\frac{1}{\left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) * (1+d_c)}}{\sum_{c=1}^{C} \frac{1}{\left(\frac{1}{n_{cA}} + \frac{1}{n_{cB}}\right) * (1+d_c)}} \qquad (5.5)$$

If a center reaches the targeted patient distribution, the distance d_c will be 0. Weighting factor II of a fixed effects model results.

5.2.3 Possibility two

Following the composition of the weighting factor of a random effects model, there is a second possibility for a new weighting factor. The center deviation will simply be added to the within-center-variability

$$\begin{aligned} w_c^{\star,V_2} &= \frac{1}{\text{within center variability + center deviation}} \\ &= \frac{1}{\sigma^2\left(\frac{1}{n_{cA}}+\frac{1}{n_{cB}}\right)+d_c} \end{aligned} \tag{5.6}$$

$$\Longrightarrow w_c^{V_2} = \frac{\frac{1}{\sigma^2\left(\frac{1}{n_{cA}}+\frac{1}{n_{cB}}\right)+d_c}}{\sum_{c=1}^{C}\frac{1}{\sigma^2\left(\frac{1}{n_{cA}}+\frac{1}{n_{cB}}\right)+d_c}} \tag{5.7}$$

If the distance of a center from the standard center is 0, this weighting factor will also reduce to weighting factor II. As the distances of all centers increase, weighting factor III is approached. The same holds for $w_c^{V_1}$. Since the maximum deviation of a center is restricted to $I * \sqrt{2}$, weighting factor III could be achieved if the number of patient characteristics I is extremely large. As previously mentioned the number of patient characteristics should not be excessively large. It then becomes almost impossible for a center to achieve the targeted sample of patients.

The difference between the new weights w_c^{V1} and w_c^{V2} (c=1,...,C) and the weights of the weight engendering function of Chapter 4 (4.6) is the fact that with regard to the new approach a center c can now have a large deviation resulting in a weighting factor which differs from w_c^{II} whereas the deviation of a second center $c^{'}$ with $c^{'} \neq c$ may be zero and will hence correspond to weight II. This was not possible for the weights of the weight engendering function $w_c^r(u)$ (4.6) for c=1,...,C. All C weights were based on the same value of u or σ_B^2, respectively.

5.3 Additional concomitant variables

Important patient characteristics are used for the definition of the standard center. Depending on the type of the patient characteristics a **covariance analysis** or an **ANOVA with additional qualitative concomitant variables** could be applied instead of the new approach. However, these methods do have a different aim. The covariance analysis as well as the ANOVA with additional qualitative concomitant variables are intended to reduce the error variance in the model, i.e. to make the analysis more precise (Neter et al. 1996 [36]). They also correct the covariate bias.

In contrast, the overall treatment effect of the new approach is biased by the treatment-by-center interaction and by the concomitant variables included in the weighting factor

$$\begin{aligned} E(\hat{\theta}^V) &= \sum_{c=1}^{C} w_c^V \left(E(\overline{Y}_{cA.}) - E(\overline{Y}_{cB.})\right) = \sum_{c=1}^{C} w_c^V (\mu_{cA} - \mu_{cB}) \\ &= \beta_A - \beta_B + \sum_{c=1}^{C} w_c^V (\alpha\beta_{cA} - \alpha\beta_{cB}) , \end{aligned}$$

where w_c^V corresponds to (5.5) or (5.7). The bias is not surprising since the aim of the new approach is to find weighting factors which are more appropriate for the centers, whereas the aim of the two types of covariate adjustment is the elimination of the covariate bias (in comparing the treatments at a common level of the covariate). The choice between the new and the former approaches should therefore depend on the aim of the analysis.

If several concomitant variables are considered as being important, it will also be possible to employ some of the covariates for the definition of the standard center and to utilize the remaining ones as supplemental covariates of the ANOVA model. Again, this should be based on the aim which is tried to achieve with the inclusion of a covariate.

Concomitant variables which are used to stratify the randomization process can also be employed for the new approach because the stratification does not influence the distributions of the concomitant variables. For example, if the randomization is stratified by "center" and "sex", patients are randomly distributed to the verum and placebo group at each center and for each sex level. The proportion of patients who are male and who are female will however not be touched.

Finally, if an ANOVA with additional qualitative concomitant variables is performed there is the problem that the model becomes very complex for more than one covariate or for interactions between the covariates and other model factors. Moreover, there might be an insufficient amount of degrees of freedom. These problems do not exist for the new approach. Its complexity does not increase remarkably when the standard center is defined by a greater number of characteristics. Furthermore, all degrees of freedom are available for the traditional model terms.

5.4 Examples

The new approach was first applied at the Transpulmin Study, a randomized controlled clinical trial conducted by 13 medical doctors in Germany between January 16, 1997 and May 7, 1998. In that study, Transpulmin balsam was tested upon its effectiveness against common colds. The trial was commissioned by Asta-Medica AG in Frankfurt. It was evaluated at the Department of Biometry and Medical Documentation of the University of Ulm. Patients between 12 and 85 years with an acute cold of the respiratory tract were included into the study. Effectiveness of Transpulmin was measured by means of an overall score which determined the degree of cold based on 6 components: weariness, lacrimation, costive nose, sore throat, coughing and dyspnea. The score ranged from 0 to 18 points, with 18 points indicating a very strong form of the cold. Primary outcome variable was the difference of the score determined at the beginning of the study and the score after 3 to 5 days after treatment. Data of the primary outcome variable of the intention-to-treat analysis were available for 128 patients of the Transpulmin group as well as for 128 patients of the placebo group (Table 5.1). The study investigated the one-sided hypothesis

H_1: Transpulmin treatment leads to a better outcome than placebo

based on a significance level of 5%.

The three characteristics "age", "smoking" and "degree of cold at the beginning of the study" were chosen retrospectively as the important patient characteristics. It was assumed that the standard center has a sample of patients with the following distributions concerning the $I = 3$ characteristics

1.) $''age''$

$$q_{X_1} = \begin{cases} 0.5 & : \quad X_1 \in [12,\ 45) \quad ''young'' \\ 0.5 & : \quad X_1 \in [45,\ > 65] \quad ''old'' \end{cases}$$

Table 5.1: Sample sizes per center and treatment of Transpulmin Study

	Number of patients at each center													
Treatment	1	2	3	4	5	6	7	8	9	10	11	12	13	$\sum$
Transpulmin	10	7	7	4	17	10	1	18	14	13	20	5	2	128
Placebo	10	9	7	3	18	10	1	17	14	10	20	5	4	128
$\sum$	20	16	14	7	35	20	2	35	28	23	40	10	6	256

2.) *"smoking"*

$$q_{X_2} = \begin{cases} 0.28 & : \quad X_2 = "smoker" \\ 0.72 & : \quad X_2 = "non-smoker" \end{cases}$$

3.) *"cold"*

$$q_{X_3} = \begin{cases} 0.5 & : \quad X_3 \in [10, 12] \quad "moderate\ cold\ at\ the\ beginning" \\ 0.5 & : \quad X_3 \in [13, 18] \quad "strong\ cold\ at\ the\ beginning" \end{cases} .$$

The first two distributions are based on data published by the Federal Statistical Office. The Federal Statistical Office announced that in 1998 there were as many 15 to 44 year-old as 45+ year-old people. For the trial under investigation it was therefore assumed that half of the study patients were up to 44 years old. Concerning the number of smokers of the German population in 1998, the Federal Statistical Office stated that out of the persons with at least 15 years 28% smoked regularly.

Patients had been included into the Transpulmin Study when they had had a degree of cold rated between 10 and the maximum of 18 at the beginning of the study. For the standard center it was expected that half of the patients have a moderate cold rated between 10 and 12 before start of treatment. Patients with a strong cold and therefore with a score between 13 and 18 points should form the other 50%.

Table 5.2: Center deviations of Transpulmin Study

	Center deviation												
Patient characteristics	1	2	3	4	5	6	7	8	9	10	11	12	13
d_{c1}	0.00	0.00	0.20	0.51	0.20	0.21	0.00	0.46	0.10	0.34	0.28	0.14	0.24
d_{c2}	0.18	0.13	0.21	0.40	0.27	0.11	0.40	0.03	0.11	0.08	0.05	0.25	0.16
d_{c3}	0.57	0.27	0.40	0.71	0.46	0.07	0.00	0.46	0.45	0.21	0.04	0.55	0.24
d_c	0.75	0.40	0.82	1.61	0.76	0.40	0.40	0.96	0.67	0.63	0.37	0.95	0.63

The Matusita distances were then calculated according to (5.3). From Table 5.2 it can be seen that the deviation of center 4 is highest. This is due to the fact that at this center all patients were non-smokers (Appendix A.2) with a moderate cold at the beginning of the study (Appendix A.3). Furthermore, most of the patients at center 4 were young (Appendix A.1). In contrast, the biggest participating center 11 has the lowest deviation when it is compared with the standard clinic.

The deviations of the different centers vary, but not in an extreme way. Hence, it can be concluded that the centers are not very heterogeneous with respect to the patient characteristics which have been chosen. The resulting weighting factors are presented in Table 5.3. For comparison purposes, weighting factors II and III have been determined as well. Since the center deviations do not vary remarkably, the new weights w^{V_1} and w^{V_2} are close to weighting factor II (± 0.3).

Table 5.3: Weighting factors of Transpulmin Study

	Center weights												
Type of weight	1	2	3	4	5	6	7	8	9	10	11	12	13
w^{III}	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
w^{II}	0.08	0.06	0.06	0.03	0.14	0.08	0.01	0.14	0.11	0.09	0.16	0.04	0.02
w^{V_1}	0.08	0.08	0.06	0.03	0.11	0.09	0.01	0.10	0.10	0.09	0.16	0.05	0.03
w^{V_2}	0.07	0.07	0.05	0.02	0.13	0.09	0.01	0.12	0.11	0.09	0.19	0.03	0.02

Data of the primary outcome variable of the Transpulmin Study were not normally distributed. Different transformations were tried. Normally distributed data

could not be achieved. Since this writing is restricted to parametric methods, normally distributed data were necessary. Another multicenter study with a normally distributed outcome was not on-hand. Hence, as a reasonable solution the median primary outcome of the Transpulmin group of each center was subtracted from the corresponding median of the placebo group. The following median differences of centers 1 to 13 resulted: +2.0, 0, -2.0, 0, +2.5, +0.5, -2.0, +4.0, +1.0, +3.5, +0.5, 0, +0.5. These differences were regarded as the center treatment effects θ_c of a fixed effects model. The estimator of the center treatment effect ($\hat{\theta}_c$) was then simulated from a normal distribution with the center treatment effect as mean and a standard deviation $v_c^2 = \sigma^2(\frac{1}{n_{cA}} + \frac{1}{n_{cB}})$ with $\sigma = 3$. σ was chosen to be 3 because this value was the estimated standard deviation of the Transpulmin Study. Results are based on 100,000 simulation runs. The expected DerSimonian and Laird moment estimate of the between-center-standard-deviation $\hat{\sigma}_B$ was determined as 1.74 and is rather small. For comparison purposes, characteristics of the traditional random approach and the random approach by Hartung as already presented in Figures 4.6 and 4.7 are repeated in Table 5.4.

Table 5.4: Characteristics of Transpulmin Study

Type of approach	$E(\hat{\theta})$	Power	p-value	Significance level
II	1.47	0.99	0.003	0.05
III	0.81	0.45	0.142	0.05
V_1	1.28	0.96	0.009	0.05
V_2	1.39	0.98	0.005	0.05
Random	1.19	0.61	0.064	0.04
Random Hartung	1.19	0.54	0.075	0.05

The expected value of the estimator of the overall treatment effect is highest for the Type II analysis and lowest for the Type III approach. The new weighting factors w_c^{V1} (5.5) and w_c^{V2} (5.7) lead to expected values between the two traditional

fixed effects approaches as do the random approaches. The expected values of the new approach are closer to the Type II value than the expected values of the random approaches. Power and p-value results of the new method are also close to the Type II approach. In fact, they are slightly worse with respect to the rejection of the nullhypothesis of no overall treatment effect. The difference is based on the heterogeneity between the centers which was quantified with the help of a standard center and included into the evaluation. The Type III approach as well as the random approaches do not have enough power to conclude a treatment effect. Overall, the new approach leads to a similar outcome than the Type II analysis because the center deviations have not been large.

In order to investigate the influence of a large deviation, a **second example** will be presented. This time data is based on a mixed effects model (σ=14, θ=10 and σ_B=1, 2, 4, 7, 10 and 20). It is assumed that 120 patients are distributed among 6 centers according to Table 4.4.

For the determination of the standard center, 10 patient characteristics are chosen. The first 5 centers are anticipated to have the desired sample of patients whereas the sixth center is assumed to deviate with the maximal possible value of $10 * \sqrt{2}$=14.14. Hence, the influence of the deviation of the sixth center on the outcome of the two-sided hypothesis for an overall treatment effect is investigated. Significance level and power are presented in Tables 5.5 and 5.6.

The results of weighting factor V_1 are in this example very close to the outcome of weight II. This is due to the composition of $w_c^{V_1}$ (5.5). The large deviation of the sixth center from the standard center is almost neutralized by the inverse of the high sample size at this center. This is not possible for weight V_2 which will therefore be favored generally. Significance level and power based on weighting factor V_2 have values between the outcome of weights II and III. As for the intermediate weighting factors of Chapter 4, this can not always be expected.

Table 5.5: Significance level of new approach based on data from a mixed effects model

	Fixed Analyses				Mixed Analyses	
Simulation	Type II	Type III	V_1	V_2	Mixed	Mixed Hartung
Mixed						
$\sigma_B = 1$	0.05	0.05	0.05	0.05	0.04	0.05
$\sigma_B = 2$	0.06	0.05	0.06	0.05	0.04	0.05
$\sigma_B = 4$	0.08	0.06	0.08	0.07	0.05	0.06
$\sigma_B = 7$	0.15	0.09	0.13	0.12	0.08	0.06
$\sigma_B = 10$	0.24	0.12	0.21	0.18	0.10	0.06
$\sigma_B = 20$	0.49	0.30	0.45	0.41	0.11	0.06

Table 5.6: Power of new approach based on data from a mixed effects model

	Fixed Analyses				Mixed Analyses	
Simulation	Type II	Type III	V_1	V_2	Mixed	Mixed Hartung
Mixed						
$\sigma_B = 1$	0.97	0.79	0.96	0.88	0.92	0.83
$\sigma_B = 2$	0.97	0.78	0.96	0.88	0.92	0.82
$\sigma_B = 4$	0.95	0.77	0.95	0.86	0.89	0.77
$\sigma_B = 7$	0.92	0.76	0.91	0.83	0.80	0.65
$\sigma_B = 10$	0.87	0.73	0.87	0.79	0.68	0.52
$\sigma_B = 20$	0.77	0.67	0.77	0.71	0.39	0.25

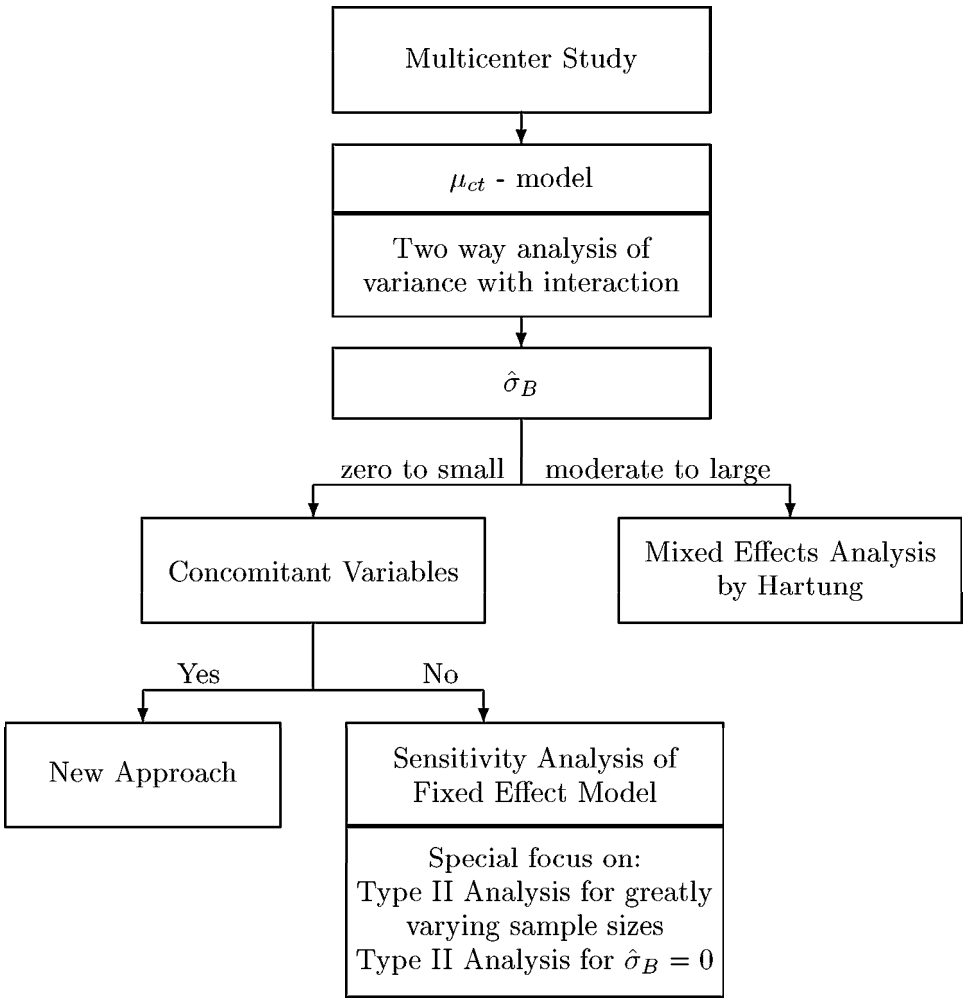

Figure 5.1: Decision on type of model

One aberrant center is therefore able to influence weighting factor V_2 so that significance level and power of the corresponding test differ from the outcome of weighting factor II. However, significance level and power of the new approach show the same behavior as the corresponding terms of the two traditional fixed effects analyses. For increases in the variance between the centers the power of analysis V_2 is higher than the power of the mixed effects approaches whereas its significance level is inflated. This is also the reason why the new approach is not able to replace the mixed effects analysis by Hartung for large interactions.

The new approach will hence be suggested if the between-center-variability has been estimated as zero to small and if important patient attributes are available (Figure 5.1). Such attributes are often on-hand. For example, concerning the field of oncology, the distribution of the start of the disease is usually known. Another example are diseases based on different causes. With the help of other important patient characteristics like e.g. gender or age, a reasonable standard center can be easily defined.

The new method has many advantages. First of all, it regards the factor "center" as fixed and offers at the same time the possibility of incorporating heterogeneity between the hospitals in the model. Thereby, the weights are considered to reflect the appropriate importance of the centers. Additionally, a decision between Analyses II and III is not necessary any more. If there are no important patient characteristics available, the proceeding of Chapter 4 will be recommended.

5.5 Future work

The main points of further interest are other types of outcome data, for example a binary outcome. Also, in case of normally distributed data, one could focus on the standardized difference in treatment means instead of using their absolute difference (Hedges and Olkin 1985 [25]). This is however more important for meta-analyses than for multicenter trials especially when different scales had been used at each site.

Next to the frequentist approaches of considering a fixed or a mixed effects model, a Bayesian approach could also be employed as was done by Jones et al. (1998 [29]) and Gould (1998 [16]). Prior distributions for the unknown parameters are required for this method.

A further point of interest could be the within-center-variability σ^2, which was assumed to be constant over all centers. If heteroscedastic variances σ_c^2 need to be anticipated for the centers or even for the treatments, the estimator of the variability changes (Hartung 1999 [20]). Classical F-tests performed under the assumption of equal error variances will provide only approximate solutions (Bao et al. 2001 [1]). The new proposed test by Hartung and Knapp (2001 [21] [22]) should be applied in this case. Finally, in order to avoid domination of a single center, Hartung and Knapp (2001 [23]) and Hartung and Makambi (2002 [24]) propose to apply a procedure which they denote as "partial weighting".

Chapter 6

Summary and Conclusions

Multicenter trials are a popular method of performing studies. In the light of globalization that is happening, multinational multicenter trials are becoming increasingly popular. One example is the opening of Japan with regard to trials performed outside the country. Japan now accepts clinical trials conducted in foreign countries on Japanese people as long as the trials are implemented according to ICH - Good Clinical Practice.

In Chapter 2 models based on fixed effects have been described in a general way. Some of the most often applied formulations for two-way analysis of variance models have been presented in order to clarify the sometimes confusing terminology which seems to differ from textbook to textbook. A unifying terminology has been suggested. This notation was then also employed for the description of Analyses I to III of fixed effects models. Characteristics derived from these methods were investigated at length. Particular attention was paid to the connection of the three analyses with terms of the Statistical Analysis System (SAS) as well as with the generalized linear model theory in order to demonstrate a unifying picture of the topic.

At the end of the chapter the opinions of various authors with regard to the three analyses were presented. It turned out that Analysis I should not be applied

generally because it is based on the assumption that a multicenter trial has been performed at a single site. However, probably due to simplicity purposes, many trials are still evaluated based on Analysis I. It coincides with Analysis II when an equal number of patients receives treatments A and B at each center. Comparing Analyses II and III, there is agreement that Analysis II will be advantageous if the sample sizes vary greatly among the centers or if there is no treatment-by-center interaction. In these cases Analysis II has more power than Analysis III to detect a treatment effect. Moreover, when the sample sizes differ greatly, Analysis III yields imprecise estimates in contrast to Analysis II. In the remaining situation, when treatment-by-center interactions are present, no consensus on the choice between Analyses II and III has been reached yet. Analysis II then leads to a biased treatment estimate. It was therefore considered as being less appropriate than Analysis III which yields an unbiased estimate. In recent times however, there are more and more statistician who prefer Analysis II not only in the absence of interactions but also when interactions are present. Their argumentation is mainly based on power considerations. New investigations concerning this topic have been performed in Chapters 4 and 5.

Chapter 3 described two-way analysis of variance models in which the centers were assumed to be a randomly chosen sample from a population of centers. Although this situation is rare in practice, the model has become an appreciated possibility of incorporating heterogeneity of treatment effects. The reason is the weighting factor of the model which does not only incorporate within-center-variability but also the variance between the centers. The different formulations of mixed effects models were also described and connections with general mixed models and with the methods available in SAS were explained. The chapter concluded by describing meta-analyses and by demonstrating that the underlying models for meta-analyses are the same as for multicenter studies.

A brief comparison of fixed and mixed approaches as well as the presentation of their theoretical peculiarities was given in Chapter 4. It was demonstrated that

the choice between a fixed and a mixed approach is still as unclear as the choice between Analysis II and Analysis III of a fixed effects model. Simulation studies were performed comparing Analysis II and Analysis III of a fixed effects model with the usual mixed effects analysis and the mixed effects analysis by Hartung. Data were therefore simulated from a fixed and from a mixed effects model.

The examination of terms related to hypothesis testing revealed that for data based on a fixed effects model, the Type II and Type III analyses often achieve a higher power for the conclusion of an overall treatment effect than the two mixed effects approaches. Since the significance levels of all methods hold the prescribed level, the fixed approaches - especially the Type II Analysis - are recommended.

If data are assumed to come from a mixed effects model, the two fixed effects analyses should not be applied when the variability between the centers is high. The reason is their inflated significance level although they have a good performance in power. The approach by Hartung is the supported method in this case because it approximately holds the prescribed level of significance. However, a low power has to be expected. On the other hand, Analysis II is proposed for a zero to small between-center-variability. It has the same significance level as the approach by Hartung but a much better performance in power. The suggestion given by Hartung and Knapp (2001 [21]) that their test should be applied in general can therefore not be supported. When the between-center-variability is of moderate magnitude and when a slight increase in the significance level is not of great concern, the traditional mixed effects model was first considered as an alternative to the approach by Hartung. This was based on the fact that a possible increase in the significance level by 0% to 2% may be more than compensated by a gain in power of 11% to 15%.

If data are based on a mixed effects model or if a fixed effects model cannot be anticipated with certainty the decision on the type of analysis should therefore be based on the magnitude of the variability between the centers. Since the variance is often not known beforehand its magnitude is commonly estimated with the method of moments by DerSimonian and Laird (1986 [7]). Unfortunately, the method does

not always yield precise estimates which might affect the choice of the approach as well as the results of both mixed approaches.

The influence of an imprecise moment estimator on the outcome of the two mixed approaches was evaluated in the following section. Furthermore, weights II and III of a fixed effects model were connected via an exponential function in order to investigate the impact of intermediate weighting factors. Hence, sensitivity analyses for both the fixed and the mixed approach were performed based on data simulated from a fixed and a mixed effects model.

The exponential connection of weights II and III of the fixed effects approach revealed for the investigated examples that corresponding significance level and power functions increased or decreased almost monotonously. There were only small rises or declines. A weighting function may however still be an additional method of exploring not only the outcome of the Type II and III analyses but also their robustness although the connection is not considered as highly important. For power reasons, special interest should be assigned to weight II in cases of greatly varying sample sizes or a zero estimate of the between-center-variability.

Concerning the sensitivity analysis of the two mixed effects models, it could be confirmed that the significance level of the Hartung approach is very robust for changes in the moment estimator of the between-center-variability. The power of the Hartung statistic proved to depend on the variance between the centers although it varied less than the power of the traditional mixed effects model. Significance level and power of the traditional mixed approach may change remarkably as the moment estimator of the between-center-variance is varied. An inflated significance level can therefore not be excluded. The high dependency of the traditional mixed approach on the estimate of the between-center-variability is the reason why this approach should not be applied when a moderate estimate is present. The Hartung approach is suggested.

Since it will often be difficult to decide if the data come from a fixed or a mixed effects model, the sensitivity analysis of a fixed effects model or the exclusive calculation of the Type II and III analyses are recommended as a first step when evaluating multicenter trials. Robustness and/or conformity of Analyses II and III can be deduced. An inflated level of significance sometimes has to be accepted due to the advantage of a much better power for finding a treatment effect. In a second step, the approach by Hartung is suggested. Lower power results have to put up with because of the excellent adherence of the approach to the prescribed level of significance. Coinciding conclusions of the fixed and mixed approaches would yield a clear outcome of the trial. Results of the Hartung approach should then be presented. If the approaches lead to different conclusions, the estimate of the between-center-variability should help to decide on the more suitable approach. More importance should be attached to the fixed analyses in case of a zero or small estimate of the between-center-variability. In all other cases, the outcome of the Hartung approach should obtain more weight.

Concerning the imprecision of the moment estimation of the between-center-variance and a possibly wrong choice of one of the approaches, a confidence interval of the between-center-variability is suggested as a means of clarifying an ambiguous situation. If a choice between the methods remains unclear and if it is of absolute importance that the test keeps the prescribed level of significance, the Hartung statistic will be favored.

Finally, in the last chapter, a new weighting factor has been introduced. It rests upon a fixed effects model. The new method incorporates heterogeneity of treatment effects by means of a "standard center" which is determined in advance of the trial. Important concomitant variables that the patients of each center should have are utilized to define the "standard sample of patients" and therefore the "standard center". The integral part of the new weighting factor is the deviation of a participating center from the "standard center". Two different possibilities of incorporating the observed deviation into a weighting factor (along with the variability within the

centers) were described. The resulting center weights were called as weight V_1 and weight V_2. Both will coincide with weighting factor II if a center recruits patients with the desired characteristics, whereas weight III will be approached if the deviations of all participating centers from the standard center are high. Weighting factor V_2 will generally be favored because a "neutralization" of an observed center deviation with a large center sample size is not possible.

In Chapter 4 it was demonstrated that significance level and power of weighting factors between weights II and III will often have a value between the outcome of weights II and III. This can however not be expected generally. The same holds for the new weight which was applied within the framework of two trials.

Since the level of significance based on the new weight is also inflated for increases in the variability between the centers it is therefore not able to replace the mixed approach by Hartung in the case of a moderate to large estimate of the between-center-variance. If the variability between the centers is estimated as zero to small and if important patient characteristics are available, weighting factor V_2 will be highly recommended. As a result, a choice between the Type II and III analyses is then not necessary any more. The main advantage of the new approach is the fact that it regards "center" as a fixed effect and simultaneously integrates heterogeneity between the hospitals into the model. Thereby, the new weighting factor is considered to reflect the appropriate importance of the centers. If there are no important patient characteristics available, the proceeding of Chapter 4 will be suggested.

Appendix A

Tables concerning the Transpulmin Study

Table A.1: Age distribution of the Transpulmin Study

	Number of patients of the two age classes	
Center	Young patients	Old patients
1	10 50.00%	10 50.00%
2	8 50.00%	8 50.00%
3	9 64.29%	5 35.71%
4	6 85.51%	1 14.20%
5	17 48.57%	18 51.43%
6	13 65.00%	7 35.00%
7	1 50.00%	1 50.00%
8	29 82.86%	6 17.14%
9	16 57.14%	12 42.86%
10	17 73.91%	6 26.09%
11	28 70.00%	12 30.00%
12	4 40.00%	6 60.00%
13	4 66.67%	2 33.33%
Standard center	50.00%	50.00%

Table A.2: Smoking distribution of the Transpulmin Study

	Number of patients of the two smoking classes	
Center	Smoker	Non-smoker
1	3	17
	15.00%	85.00%
2	3	13
	18.75%	81.25%
3	6	8
	42.86%	57.14%
4	0	7
	0.00%	100.00%
5	3	32
	8.57%	91.43%
6	4	16
	20.00%	80.00%
7	0	2
	0.00%	100.00%
8	9	26
	25.71%	74.29%
9	9	16
	36.00%	64.00%
10	7	14
	33.33%	66.67%
11	9	28
	24.32%	75.68%
12	1	9
	10.00%	90.00%
13	1	5
	16.67%	83.33%
Standard center	28.00%	72.00%

Table A.3: Distribution of degree of cold at the beginning of the Transpulmin Study

	Number of patients of the two cold classes	
Center	Moderate cold	Strong cold
1	18 90.00%	12 10.00%
2	11 68.75%	5 31.25%
3	11 78.57%	3 21.43%
4	7 100.00%	0 0.00%
5	29 82.86%	6 17.14%
6	11 55.00%	9 45.00%
7	1 50.00%	1 50.00%
8	29 82.86%	6 17.14%
9	23 82.14%	5 17.86%
10	15 65.22%	8 34.78%
11	21 52.50%	19 47.50%
12	8 88.89%	1 11.11%
13	4 66.67%	2 33.33%
Standard center	50.00%	50.00%

Bibliography

[1] P. Bao and M.M.A. Ananda (2001). "Performance of two-way anova procedures when cell frequencies and variances are unequal". *Communications in Statistics*, Vol. 30, No. 4, pp. 805-829.

[2] D.A. Berry (1990). "Basic Principles in Designing and Analyzing Clinical Studies". In: D.A. Berry (ed.), *Statistical Methodology in the Pharmaceutical Sciences*, Marcel Dekker 1990, Chapter 7.

[3] β-Blocker Heart Attack Trial Research Group (1982). "A randomized trial of propanolol in patients with acute myocardial infarction. I. Mortality Results". *JAMA*, Vol. 247, No. 12, pp. 1707-1714.

[4] B.J. Biggerstaff and R.L. Tweedie (1997). "Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis". *Statistics in Medicine*, Vol. 16, pp. 753-768.

[5] W.G. Cochran (1954). "The combination of estimates from different experiments". *Biometrics*, Vol. 10, pp. 101-129.

[6] W.G. Cochran (1983). "Adjustments in analysis". In *Planning and Analysis of Observational Studies*, Wiley, New York, pp. 102-108.

[7] R. DerSimonian and N.M. Laird (1986). "Meta-analysis in clinical trials". *Controlled Clinical Trials*, Vol. 7, pp. 177-188.

[8] A. Dobson (1990). *An introduction to generalized linear models.* Chapmann and Hall.

[9] R.C. Elston and N. Bush (1964). "The hypotheses can be tested when there are interactions in an analysis of variance model". *Biometrics*,Vol. 20, pp. 681-698.

[10] J.L. Fleiss (1985). "Multicenter studies". In: *The Design and Analysis of Clinical Experiments*, Wiley, pp. 176-185.

[11] J.L. Fleiss (1986). "Analysis of data from multiclinic trials". *Controlled Clinical Trials*, Vol. 7, pp. 267-275.

[12] P.R. Freeman (1989). "The performance of the two-stage analysis of two-treatment, two-period cross-over trials". *Statistics in Medicine*, Vol. 8, pp. 1421-1432.

[13] M. Gail and R. Simon (1985). "Testing for qualitative interactions between treatment effects and patient subsets". *Biometrics*, Vol. 41, pp. 361-372.

[14] P. Gallo (2000). "Center-weighting issues in multicenter clinical trials". *Journal of Biopharmaceutical Statistics*, Vol. 10, No. 2, pp. 145-163.

[15] J.D. Goldberg and K.J. Koury (1990). "Design and analysis of multicenter trials". In: D.A. Berry (ed.), *Statistical Methodology in the Pharmaceutical Sciences*, Marcel Dekker, Chapter 7.

[16] A.L. Gould (1998). "Multi-centre trial analysis revisited". *Statistics in Medicine*, Vol. 17, pp. 1779-1797.

[17] F.A. Graybill (1976). *Theory and application of the linear model.* North Scituate, Mass. Duxbury Pr..

[18] R.J. Hardy, S.G. Thompson (1996). "A likelihood approach to meta-analysis with random effects". *Statistics in Medicine*, Vol. 15, pp. 619-629.

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

[20] J. Hartung (1999). "An alternative method for meta-analysis". *Biometrical Journal*, Vol. 41, No. 8, pp. 901-916.

[21] J. Hartung and G. Knapp (2001). "On tests of the overall treatment effect in meta-analysis with normally distributed responses". *Statistics in Medicine*, Vol. 20, pp. 1771-1782.

[22] J. Hartung and G. Knapp (2001). "A redefined method for the meta-analysis of controlled clinical trials with binary outcome". *Statistics in Medicine*, Vol. 20, pp. 3875-3889.

[23] J. Hartung and G. Knapp (2001). "Models in combining of results: retrospective and prospective". *Manuscript of the XXIst ISCB Conference 2002 in Freiburg.*

[24] J. Hartung and K.H. Makambi (2002). "Positive estimation of the between-study-variance in meta-analysis". *South African Statistical Journal*, Vol. 36, pp. 57-78.

[25] L.V. Hedges and I. Olkin (1985). *Statistical methods for meta-analysis.* Academic Presss, Orlando.

[26] R.I. Horwitz, B.H. Singer, R.W. Makuch and C.M. Viscoli (1996). "Can treatment that is helpful on average be harmful to some patients ? A study of the conflicting information needs of clinical inquiry and drug regulation". *Journal of Clinical Epidemiology*, Vol. 49, No. 4, pp. 395-400.

[27] H. Hsin-Tsung, C. Shein-Chung (1998). "Design and analysis of multinational clinical trials". *Drug Information Journal*, Vol. 32, pp. 1309S-1316S.

[28] U. Jensen (1993). *Herleitung, Berechnung und ökonomische Anwendung von Rao-Distanzen.* Reihe: Quantitative Ökonomie, Band 40, Verlag Josef Eul, Bergisch Gladbach.

[29] B. Jones, D. Teather, J. Wang and J.A. Lewis (1998). "A comparison of various estimators of a treatment difference for a multi-centre clinical trial". *Statistics in Medicine*, Vol. 17, pp. 1767-1777.

[30] J. Kaufmann (2001). "Analyse multizentrischer Studien - Parametrische Methoden". *46. GMDS Jahrestagung*, Schering AG Berlin, Diagnostika-Biometrie.

[31] A. Källén (1997). "Treatment-by-center interaction: what is the issue?" *Drug Information Journal*, Vol. 31, pp. 927-936.

[32] S. Kullback (1959). *Information theory and statistics.* Dover Publications, 1997.

[33] S. Kullback and R.A. Leibler (1951). "On information and sufficiency". *Annals of Mathematical Statistics*, Vol. 22, pp. 79-86.

[34] Z. Lin (1999). "An Issue of Statistical analysis in controlled multi-centre studies: How shall we weight the centres?" *Statistics in Medicine*, Vol. 18, pp. 365-373.

[35] K. Matusita (1955). "Decision rules, based on the distance, for problems of fit, two samples and estimation". *Annals of Mathematical Statistics*, Vol. 26, pp. 631-640.

[36] J. Neter, H. Kutner, C.J. Nachtsheim and W. Wassermann (1996). *Applied Linear Statistical Models.* McGraw-Hill Companies, Fourth Edition.

[37] A. Phillips, A. Ebbutt, L. France, D. Morgan (2000). "The international conference on harmonization guideline; "Statistical principles for clinical trials":Issues in applying the guideline in practice". *Drug Information Journal*, Vol. 34, pp. 337-348.

[38] D. Salsburg (1999). "Why analysis of variance is inappropriate for multiclinic trials". *Controlled clinical trials*, Vol. 20, pp. 453-468.

[39] SAS Institute Inc. (1991). SAS system for linear models. SAS series in statistical applications.

[40] SAS Institute Inc. Copyright (c) 1999. SAS OnlineDoc, Version 8. Cary, NC, USA.

[41] H. Scheffé (1964). *The Analysis of Variance.* John Wiley and Sons, New York.

[42] S.R. Searle (1971). *Linear Models.* Wiley.

[43] S.R. Searle, F.M. Speed and H.V. Henderson (1981). "Some computational and model equivalences in analysis of variance of unequal-subclass-numbers data". *The American Statistician*, Vol. 35, No. 1.

[44] S.S. Searle (1987). *Linear models for unbalanced data.* John Wiley & Sons.

[45] S.S. Searle, G. Casella and C.E. McCulloch (1992). *Variance components.* John Wiley & Sons.

[46] S. Senn (1997). *Statistical Issues in drug development.* Chichester, UK: Wiley.

[47] S. Senn and F. Harrell (1997). "On wisdom after the event". *Journal of Clinical Epidemiology*, Vol. 50, Nr. 7, pp. 749-751.

[48] S. Senn (1998). "Some controversies in planning and analysing multi-centre studies". *Statistics in Medicine*, Vol. 17, pp. 1753-1765.

[49] S. Senn (2000). "The many modes of meta". *Drug information journal*, Vol. 34, pp. 535-549.

[50] C.E. Shannon (1948). "A mathematical theory of communication". *Bell System Technical Journal*, Vol. 27, pp. 379-423 and 623-656.

[51] S.G. Thompson and S.J. Sharp (1999). "Explaining heterogeneity in meta-analysis: a comparison of methods". *Statistics in Medicine*, Vol. 18, pp. 2693-2708.

[52] J. Vilar, M.E. Mackey, G. Carroli and A. Donner (2001). "Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: com-

parison of fixed and random effects models". *Statistics in Medicine*, Vol. 20, pp. 3635-3647.

[53] A. Whitehead and J. Whitehead (1991). "A general parametric approach to the meta-analysis of randomized clinical trials". *Statistics in Medicine*, Vol. 10, pp. 1665-1677.

[54] F. Yates (1934). "The analysis of multiple classifications with unequal numbers in the different classes". *Journal of the American Statistical Association*, Vol. 29, pp. 51-66.

Acknowledgements

First of all I would like to thank PD Dr. Josef Högel for his helpful advice and assistance with regard to this work. Furthermore, I am also very grateful for the support of Prof. Dr. Wilhelm Gaus and of all my colleagues of the Department of Biometry and Medical Documentation of the University of Ulm.

Last but not least, I would like to thank Viatris (the former Asta Medica) in Frankfurt for their kind permission to use data from the Transpulmin Trial.