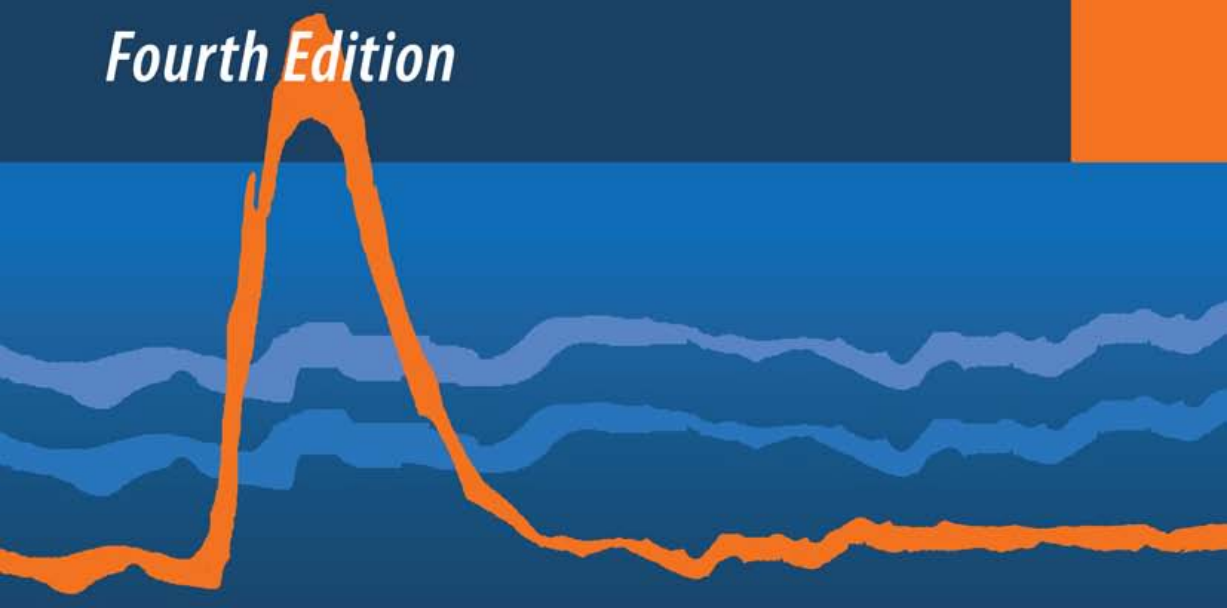
Ton J. Cleophas
Aeilko H. Zwinderman
Toine F. Cleophas
Eugene P. Cleophas

Statistics Applied to Clinical Trials

Fourth Edition



Springer

STATISTICS APPLIED TO CLINICAL TRIALS

FOURTH EDITION

Statistics Applied to Clinical Trials

Fourth edition

by

TON J. CLEOPHAS, MD, PhD, Professor
Statistical Consultant, Circulation, Boston, USA,
Co-Chair Module Statistics Applied to Clinical Trials,
European Interuniversity College of Pharmaceutical Medicine, Lyon, France,
Internist-clinical pharmacologist,
Department Medicine, Albert Schweitzer Hospital, Dordrecht, The Netherlands

AEILKO H. ZWINDERMAN, MathD, PhD, Professor
Co-Chair Module Statistics Applied to Clinical Trials,
European Interuniversity College of Pharmaceutical Medicine, Lyon, France,
Professor of Statistics,
Department Biostatistics and Epidemiology, Academic Medical Center, Amsterdam,
The Netherlands

TOINE F. CLEOPHAS, BSc
Department of Research, Damen Shipyards, Gorinchem, The Netherlands

and

EUGENE P. CLEOPHAS, BSc
Technical University, Delft, The Netherlands

Springer

Prof. Ton J. Cleophas
Albert Schweitzer Hospital
Dordrecht
The Netherlands

Prof. Aeilko H. Zwinderman
Academic Medical Center
Amsterdam
The Netherlands

Toine F. Cleophas
Damen Shipyards
Gorinchem
The Netherlands

Eugene P. Cleophas
Technical University
Delft
The Netherlands

ISBN 978-1-4020-9522-1 e-ISBN 978-1-4020-9523-8

Library of Congress Control Number: 2008939866

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

TABLE OF CONTENTS

PREFACE TO FIRST EDITION

The European Interuniversity Diploma of Pharmaceutical Medicine is a postacademic course of 2-3 years sponsored by the Socrates program of the European Community. The office of this interuniversity project is in Lyon and the lectures are given there. The European Community has provided a building and will remunerate lecturers. The institute which provides the teaching is called the European College of Pharmaceutical Medicine, and is affiliated with 15 universities throughout Europe, whose representatives constitute the academic committee. This committee supervises educational objectives. Start lectures February 2000.

There are about 20 modules for the first two years of training, most of which are concerned with typically pharmacological and clinical pharmacological matters including pharmacokinetics, pharmacodynamics, phase III clinical trials, reporting, communication, ethics and, any other aspects of drug development. Subsequent training consists of practice training within clinical research organisations, universities, regulatory bodies etc., and finally of a dissertation. The diploma, and degree are delivered by the Claude Bernard University in Lyon as well as the other participating universities.

The module "Statistics applied to clinical trials" will be taught in the form of a 3 to 6 day yearly course given in Lyon and starting February 2000. Lecturers have to submit a document of the course (this material will be made available to students). Three or 4 lecturers are requested to prepare detailed written material for students as well as to prepare examination of the students. The module is thus an important part of a postgraduate course for physicians and pharmacists for the purpose of obtaining the European diploma of pharmaceutical medicine. The diploma should make for leading positions in pharmaceutical industry, academic drug research, as well as regulatory bodies within the EC. This module is mainly involved in the statistics of randomized clinical trials.

The chapters 1-9, 11, 17, 18 of this book are based on the module "Medical statistics applied to clinical trials" and contain material that should be mastered by the students before their exams. The remaining chapters are capita selecta intended for excellent students and are not included in the exams.

The authors believe that this book is innovative in the statistical literature because, unlike most introductory books in medical statistics, it provides an explanatory rather than mathematical approach to statistics, and, in addition, emphasizes non-classical but increasingly frequently used methods for the statistical analyses of clinical trials, e.g., equivalence testing, sequential analyses, multiple linear regression analyses for confounding, interaction, and synergism. The authors are not aware of any other work published so far that is comparable with the current work, and, therefore, believe that it does fill a need.

August 1999
Dordrecht, Leiden , Delft

PREFACE TO SECOND EDITION

In this second edition the authors have removed textual errors from the first edition. Also seven new chapters (chapters 8, 10, 13, 15-18) have been added. The principles of regression analysis and its resemblance to analysis of variance was missing in the first edition, and have been described in chapter 8. Chapter 10 assesses curvilinear regression. Chapter 13 describes the statistical analyses of crossover data with binary response. The latest developments including statistical analyses of genetic data and quality-of-life data have been described in chapters 15 and 16. Emphasis is given in chapters 17 and 18 to the limitations of statistics to assess non-normal data, and to the similarities between commonly-used statistical tests. Finally, additional tables including the Mann-Whitney and Wilcoxon rank sum tables have been added in the Appendix.

December 2001, Dordrecht, Amsterdam, Delft

PREFACE TO THE THIRD EDITION

The previous two editions of this book, rather than having been comprehensive, concentrated on the most relevant aspects of statistical analysis. Although well-received by students, clinicians, and researchers, these editions did not answer all of their questions. This called for a third, more comprehensive, rewrite. In this third edition the 18 chapters from the previous edition have been revised, updated, and provided with a conclusions section summarizing the main points. The formulas have been re-edited using the Formula-Editor from Windows XP 2004 for enhanced clarity. Thirteen new chapters (chapters 8-10, 14,15, 17, 21, 25-29, 31) have been added. The chapters 8-10 give methods to assess the problems of multiple testing and data testing closer to expectation than compatible with random. The chapters 14 and 15 review regression models using an exponential rather than linear relationship including logistic, Cox, and Markow models. Chapter 17 reviews important interaction effects in clinical trials and provides methods for their analysis. In chapter 21 study designs appropriate for medicines from one class are discussed. The chapters 25-29 review respectively (1) methods to evaluate the presence of randomness in the data, (2) methods to assess variabilities in the data, (3) methods to test reproducibility in the data, (4) methods to assess accuracy of diagnostic tests, and (5) methods to assess random rather than fixed treatment effects. Finally, chapter 31 reviews methods to minimize the dilemma between sponsored research and scientific independence. This updated and extended edition has been written to serve as a more complete guide and reference-text to students, physicians, and investigators, and, at the same time, preserves the common sense approach to statistical problem-solving of the previous editions.

August 2005, Dordrecht, Amsterdam, Delft

PREFACE TO FOURTH EDITION

In the past few years many important novel methods have been applied in published clinical research. This has made the book again rather incomplete after its previous edition. The current edition consists of 16 new chapters, and updates of the 31 chapters from the previous edition. Important methods like Laplace transformations, log likelihood ratio statistics, Monte Carlo methods, and trend testing have been included. Also novel methods like superiority testing, pseudo-R2 statistics, optimism corrected c-statistic, I-statistics, and diagnostic meta-analyses have been addressed.
The authors have given special efforts for all chapters to have their own introduction, discussion, and references section. They can, therefore, be studied separately and without need to read the previous chapters first.

September 2008, Dordrecht, Amsterdam, Gorinchem, and Delft

FOREWORD

In clinical medicine appropriate statistics has become indispensable to evaluate treatment effects. Randomized controlled trials are currently the only trials that truly provide evidence-based medicine. Evidence based medicine has become crucial to optimal treatment of patients. We can define randomized controlled trials by using Christopher J. Bulpitt's definition "a carefully and ethically designed experiment which includes the provision of adequate and appropriate controls by a process of randomization, so that precisely framed questions can be answered". The answers given by randomized controlled trials constitute at present the way how patients should be clinically managed. In the setup of such randomized trial one of the most important issues is the statistical basis. The randomized trial will never work when the statistical grounds and analyses have not been clearly defined beforehand. All endpoints should be clearly defined in order to perform appropriate power calculations. Based on these power calculations the exact number of available patients can be calculated in order to have a sufficient quantity of individuals to have the predefined questions answered. Therefore, every clinical physician should be capable to understand the statistical basis of well performed clinical trials. It is therefore a great pleasure that Drs. T.J. Cleophas, A.H. Zwinderman, and T.F. Cleophas have published a book on statistical analysis of clinical trials. The book entitled "Statistics Applied to Clinical Trials" is clearly written and makes complex issues in statistical analysis transparant. Apart from providing the classical issues in statistical analysis, the authors also address novel issues such as interim analyses, sequential analyses, and meta-analyses. The book is composed of 18 chapters, which are nicely structured. The authors have deepened our insight in the applications of statistical analysis of clinical trials. We would like to congratulate the editors on this achievement and hope that many readers will enjoy reading this intriguing book.

E.E. van der Wall, MD, PhD, Professor of Cardiology, President Netherlands Association of Cardiology, Leiden, The Netherlands

CHAPTER 1

HYPOTHESES, DATA, STRATIFICATION

1. GENERAL CONSIDERATIONS

Over the past decades the randomized clinical trial has entered an era of continuous improvement and has gradually become accepted as the most effective way of determining the relative efficacy and toxicity of new drug therapies. This book is mainly involved in the methods of prospective randomized clinical trials of new drugs. Other methods for assessment including open-evaluation-studies, cohort- and case-control studies, although sometimes used, e.g., for pilot studies and for the evaluation of long term drug-effects, are excluded in this course. Traditionally, clinical drug trials are divided into IV phases (from phase I for initial testing to phase IV after release for general use), but scientific rules governing different phases are very much the same, and can thus be discussed simultaneously.

A. CLEARLY DEFINED HYPOTHESES

Hypotheses must be tested prospectively with hard data, and against placebo or known forms of therapies that are in place and considered to be effective. Uncontrolled studies won't succeed to give a definitive answer if they are ever so clever. Uncontrolled studies while of value in the absence of scientific controlled studies, their conclusions represent merely suggestions and hypotheses. The scientific method requires to look at some controls to characterize the defined population.

B. VALID DESIGNS

Any research but certainly industrially sponsored drug research where sponsors benefit from favorable results, benefits from valid designs. A valid study means a study unlikely to be biased, or unlikely to include systematic errors. The most dangerous errors in clinical trials are systematic errors otherwise called biases. Validity is the most important thing for doers of clinical trials to check. Trials should be made independent, objective, balanced, blinded, controlled, with objective measurements, with adequate sample sizes to test the expected treatment effects, with random assignment of patients.

C. EXPLICIT DESCRIPTION OF METHODS

Explicit description of the methods should include description of the recruitment procedures, method of randomization of the patients, prior statements about the methods of assessments of generating and analysis of the data and the statistical methods used, accurate ethics including written informed consent.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 1–16.

D. UNIFORM DATA ANALYSIS

Uniform and appropriate data analysis generally starts with plots or tables of actual data. Statistics then comes in to test primary hypotheses primarily. Data that do not answer prior hypotheses may be tested for robustness or sensitivity, otherwise called precision of point estimates e.g., dependent upon numbers of outliers. The results of studies with many outliers and thus little precision should be interpreted with caution. It is common practice for studies to test multiple measurements for the purpose of answering one single question. In clinical trials the benefit to health is estimated by variables, which can be defined as measurable factors or characteristics used to estimate morbidity / mortality / time to events etc. Variables are named exposure, indicator, or independent variables, if they predict morbidity / mortality, and outcome or dependent variables, if they estimate morbidity / mortality. Sometimes both mortality and morbidity variables are used in a single trial, and there is nothing wrong with that practice. We should not make any formal correction for multiple comparisons of this kind of data. Instead, we should informally integrate all the data before reaching conclusions, and look for the trends without judging one or two low P-values among otherwise high P-values as proof.

However, subgroup analyses involving post-hoc comparisons by dividing the data into groups with different ages, prior conditions, gender etc can easily generate hundreds of P-values. If investigators test many different hypotheses, they are apt to find significant differences at least 5% of the time. To make sense of these kinds of results, we need to consider the Bonferroni inequality, which will be emphasized in the chapters 7 and 8. It states that, if k statistical tests are performed with the cut-off level for a test statistic, for example t or F, at the α level, the likelihood for observing a value of the test statistic exceeding the cut-off level is no greater than k times α. For example, if we wish to do three comparisons with t-tests while keeping the probability of making a mistake less than 5%, we have to use instead of α =5% in this case α =5/3%=1.6%. With many more tests, analyses soon lose any sensitivity and do hardly prove anything anymore. Nonetheless, a limited number of post-hoc analyses, particularly if a plausible theory is underlying, can be useful in generating hypotheses for future studies.

2. TWO MAIN HYPOTHESES IN DRUG TRIALS: EFFICACY AND SAFETY

Drug trials are mainly for addressing the efficacy as well as the safety of the drugs to be tested in them. For analyzing efficacy data formal statistical techniques are normally used. Basically, the null hypothesis of no treatment effect is tested, and is rejected when difference from zero is significant. For such purpose a great variety of statistical significance tests has been developed, all of whom report P values, and compute confidence intervals to estimate the magnitude of the treatment effect. The appropriate test depends upon the type of data and will be discussed in the next chapter. Of safety data, such as adverse events, data are mostly collected with

the hope of demonstrating that the test treatment is not different from control. This concept is based upon a different hypothesis from that proposed for efficacy data, where the very objective is generally to show that there actually is a difference between test and control. Because the objective of collecting safety data is thus different, the approach to analysis must be likewise different. In particular, it may be less appropriate to use statistical significance tests to analyze the latter data. A significance test is a tool that can help to establish whether a difference between treatments is likely to be real. It cannot be used to demonstrate that two treatments are similar in their effects. In addition, safety data, more frequently than efficacy data, consist of proportions and percentages rather than continuous data as will be discussed in the next section. Usually, the best approach to analysis of these kinds of data is to present suitable summary statistics, together with confidence intervals. In the case of adverse event data, the rate of occurrence of each distinct adverse event on each treatment group should be reported, together with confidence intervals for the difference between the rates of occurrence on the different treatments. An alternative would be to present risk ratios or relative risks of occurrence, with confidence intervals for the relative risk. Chapter 3 mainly addresses the analyses of these kinds of data.
Other aspects of assessing similarity rather than difference between treatments will be discussed separately in chapter 6 where the theory, equations, and assessments are given for demonstrating statistical equivalence.

3. DIFFERENT TYPES OF DATA: CONTINUOUS DATA

The first step, before any analysis or plotting of data can be performed, is to decide what kind of data we have. Usually data are continuous, e.g., blood pressures, heart rates etc. But, regularly, proportions or percentages are used for the assessment of part of the data. The next few lines will address how we can summarize and characterize these two different approaches to the data.

Samples of **continuous data** are characterized by:

$$\textbf{Mean} = \frac{\Sigma x}{n} = \bar{x},$$

where Σ is the summation, x are the individual data, n is the total number of data.

$$\textbf{Variance between the data} = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$\textbf{Standard deviation (SD)} = \sqrt{(Variance)}$$

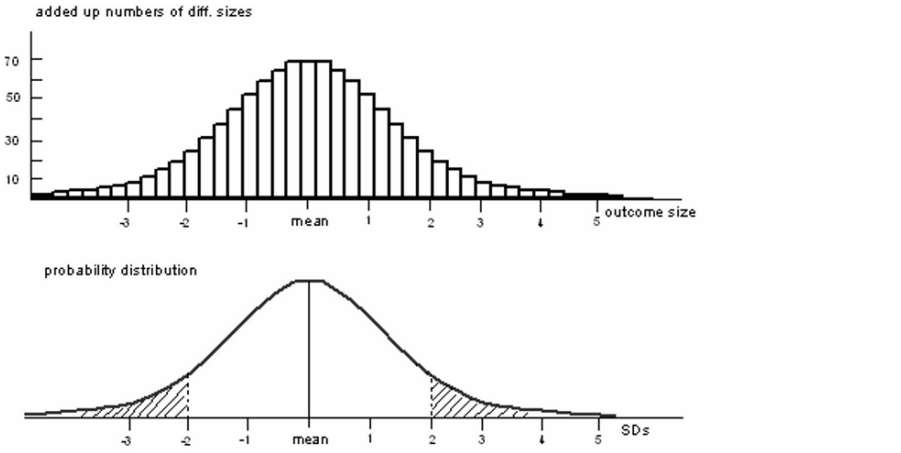


Figure 1. Histogram and Gaussian curve representation of data.

Continuous data can be plotted in the form of a histogram (Figure 1 upper graph). On the x-axis, frequently called z-axis in statistics, it has individual data. On the y-axis it has "how often". For example, the mean value is observed most frequently, while the bars on either side gradually grow shorter. This graph adequately represents the data. It is, however, not adequate for statistical analyses. Figure 1 lower graph pictures a Gaussian curve, otherwise called normal (distribution) curve. On the x-axis we have, again, the individual data, expressed either in absolute data or in SDs distant from the mean. On the y-axis the bars have been replaced with a continuous line. It is now impossible to determine from the graph how many patients had a particular outcome. Instead, important inferences can be made. For example, the total area under the curve (AUC) represents 100% of the data, AUC left from mean represents 50% of the data, left from -1 SDs it has 15.87% of the data, left from -2SDs it has 2.5% of the data. This graph is better for statistical purposes but not yet good enough.

Figure 2 gives two Gaussian curves, a narrow and a wide one. Both are based on

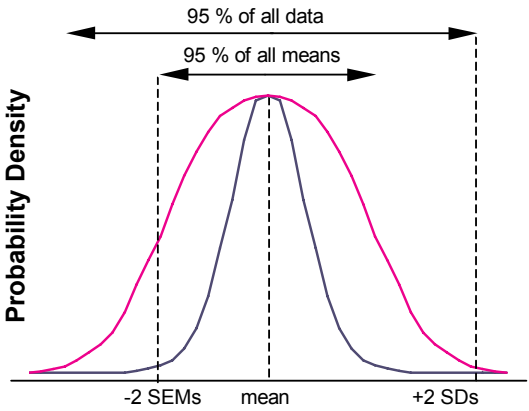


Figure 2. Two examples of normal distributions.

the same data, but with different meaning. The wide one summarizes the data of our trial. The narrow one summarizes the mean of many trials similar to our trial. We will not try to make you understand why this is so. Still, it is easy to conceive that the distribution of all means of many similar trials is narrower and has fewer outliers than the distribution of the actual data from our trial, and that it will center (centre) around the mean of our trial because our trial is assumed to be representative for the entire population. You may find it hard to believe, but the narrow curve with standard errors of the mean (SEMs) or simply SEs on the x-axis can be effectively used for testing important statistical hypotheses, like (1) no difference between new and standard treatment, (2) a real difference, (3) the new treatment is better than the standard treatment, (4) the two treatments are equivalent. Thus, mean ± 2 SDs (or more precisely 1.96 SDs) represents the AUC of the wide distribution, otherwise called the 95% confidence interval of the data, which means that 95 % of the data of the sample are within. The SEM-curve (narrow one) is narrower than the SD-curve (wide one) because SEM = SD/$\sqrt{n}$ with n = sample size. Mean ± 2 SEMs (or more precisely 1.96 SEMs) represents 95% of the means of many trials similar to our trial.

$$\mathbf{SEM = SD / \sqrt{n}}$$

As the size of SEM in the graph is about 1/3 times SD, the size of each sample is here about n = 10. The area under the narrow curve represents 100 % of the sample means we would obtain, while the area under the curve of the wide graph represents 100% of all of the data of the samples.

Why is this SEM approach so important in statistics. Statistics makes use of mean values and their standard error to test the null hypotheses of finding no difference

from zero in your sample. When we reject a null hypothesis at P<0.05, it literally means that there is < 5 % chance that the mean value of our sample crosses the area of the null hypothesis where we say there is no difference. It does not mean that many individual data may not go beyond that boundary. Actually, it is just a matter of agreement. But it works well.

So remember:
Mean ± 2 SDs covers an area under the curve including 95 % of the data of the given sample.
Mean ± 2 SEMs covers an area under curve including 95% of the means of many samples, and is sometimes called the 95% confidence interval (CI).

In statistical analysis we often compare different samples by taking their sums or differences. Again, this text is not intended to explain the procedures entirely. One more thing to accept unexplainedly is the following. The distributions of the sums as well as those of the difference of samples are again normal distributions and can be characterized by:

$$\text{Sum: } \text{mean}_1 + \text{mean}_2 \pm \sqrt{(\text{SD}_1^2 + \text{SD}_2^2)}$$
$$\text{Difference: } \text{mean}_1 - \text{mean}_2 \pm \sqrt{(\text{SD}_1^2 + \text{SD}_2^2)}$$

$$\text{SEM}_{\text{sum}} = \sqrt{(\text{SD}_1^2/\text{n}_1 + \text{SD}_2^2/\text{n}_2)}$$

$\text{SEM}_{\text{difference}}$= “

Note: If the standard deviations are very different in size, then a more adequate calculation of the pooled SEM is given in the next chapter.
Sometimes we have paired data where two experiments are performed in one subject or in two members of one family. The variances with paired data are usually smaller than with unpaired because of the positive correlation between two observations in one subject (those who respond well the first time are more likely to do so the second). This phenomenon translates in a slightly modified calculation of variance parameters.

$$\text{SD}_{\text{paired sum}} = \sqrt{(\text{SD}_1^2 + \text{SD}_2^2 + 2\,\text{r}\,\text{SD}_1 \cdot \text{SD}_2)}$$
$$\text{SD}_{\text{paired differrence}} = \sqrt{(\text{SD}_1^2 + \text{SD}_2^2 - 2\,\text{r}\,\text{SD}_1 \cdot \text{SD}_2)}$$

Where r = correlation coefficient, a term that will be explained soon.

Likewise:

$$SEM_{paired\,sum} = \sqrt{SD_1^2/n_1 + SD_2^2/n_2 + (2\,r\,SD_1 \cdot SD_2)\,(1/2n_1 + 1/2n_2)}$$

$$SEM_{paired\ differrence} = \sqrt{SD_1^2/n_1 + SD_2^2/n_2 - (2\,r\,SD_1 \cdot SD_2)\,(1/2n_1 + 1/2n_2)}$$

Note that SEM does not directly quantify variability in a population. A small SEM can be mainly due to a large sample size rather than tight data.
With small samples the distribution of the means does not exactly follow a Gaussian distribution. But rather a t-distribution, 95% confidence intervals cannot be characterized as the area under the curve between mean ± 2 SEMs but instead the area under curve is substantially wider and is characterized as mean ± t.SEMs where t is close to 2 with large samples but 2.5-3 with samples as small as 5-10. The appropriate t for any sample size is given in the t-table.

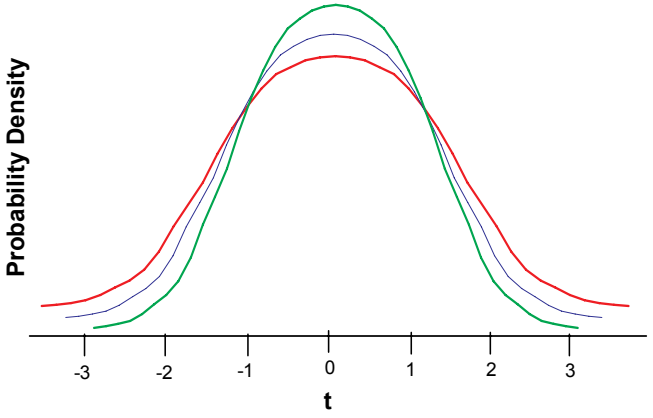


Figure 3. Family of t-distributions: with n=5 the distribution is wide, with n=10 and n=1000 this is increasingly less so.

Figure 3 shows that the t-distribution is wider than the Gaussian distribution with small samples. Mean ± t.SEMs presents the 95 % confidence intervals of the means that many similar samples would produce.

Statistics is frequently used to compare more than 2 samples of data. To estimate whether differences between samples are true or just chance we first assess variances in the data between groups and within groups.

Group	n patients	mean	SD
Group 1	n	mean $_1$	SD_1
Group 2	n	mean $_2$	SD_2
Group 3	n	mean $_3$	SD_3

This procedure may seem somewhat awkward in the beginning but in the next two chapters we will observe that variances, which are no less than estimates of noise in the data, are effectively used to test the probabilities of true differences between, e.g., different pharmaceutical compounds. The above data are summarized underneath.

Between-group variance:

Sum of squares$_{between}$ = $SS_{between}$ = n (mean$_1$ – overall mean)^2 + n (mean$_2$ – overall mean)^2 + n (mean$_3$ – overall mean)^2

Within-group variance:

Sum of squares$_{within}$ = SS_{within} = (n-1) SD_1^2 + (n-1) SD_2^2 + (n-1) SD_3^2

The ratio of the sum of squares between-group / sum of squares within group (after proper adjustment for the sample sizes or degrees of freedom, a term which will be explained later on) is called the big F and determines whether variances between the sample means is larger than expected from the variability within the samples. If so, we reject the null hypothesis of no difference between the samples. With two samples the square root of big F, which actually is the test statistic of analysis of variance (ANOVA), is equal to the t of the famous t-test, which will further be explained in chapter 2. These 10 or so lines already brought us very close to what is currently considered the heart of statistics, namely ANOVA (analysis of variance).

4. DIFFERENT TYPES OF DATA: PROPORTIONS, PERCENTAGES AND CONTINGENCY TABLES

Instead of continuous data, data may also be of a discrete character where two or more alternatives are possible, and, generally, the frequencies of occurrence of each of these possibilities are calculated. The simplest and commonest type of such data are the binary data (yes/no etc). Such data are frequently assessed as proportions or percentages, and follow a so-called binomial distribution. If 0.1< proportion (p) <0.9 the binomial distribution becomes very close to the normal distribution. If p <0.1, the data will follow a skewed distribution, otherwise

called Poisson distribution. Proportional data can be conveniently laid-out as contingency tables. The simplest contingency table looks like this:

	numbers of subjects with side Effect	numbers of subjects without side effect
Test treatment ($group_1$)	a	b
Control treatment ($group_2$)	c	d

The proportion of subjects who had a side effect in group (or the risk (**R**) or probability of having an effect):

p= a / (a+b) , in $group_2$ **p= c / (c+d)**,

The ratios **a / (a+b)** and **c / (c+d)** are called **risk ratios (RRs)**

Note that the terms proportion, risk and probability are frequently used in statistical procedures but that they basically mean the same.

Another approach is the **odds** approach **a/b** and **c/d** are odds and their ratio is the **odds ratio (OR)**.
In clinical trials we use ORs as surrogate RRs, because here a/(a+b) is simply nonsense. For example:

	treatment-group		control-group		entire-population
sleepiness	32	a	4	b	4000
no sleepiness	24	c	52	d	52000

We assume that the control group is just a sample from the entire population but that the ratio b/d is that of the entire population. So, suppose 4=4000 and 52= 52000, then we can approximate $\frac{a/(a+b)}{c/(c+d)} = \frac{a/b}{c/d}$ = RR of the entire population.
With observational cohort studies things are different. The entire population is used as control group. Therefore, RRs are better adequate. Ors and RRs are largely similar as long as they are close to 1.000. More information on Ors is given in the Chapters 3, 16, and 44.

Proportions can also be expressed as percentages:

p.100 %= a / (a+b). (100%) etc

Just as with continuous data we can calculate SDs and SEMs and 95% confidence intervals of rates (or numbers, or scores) and of proportions or percentages.

$$\textbf{SD} \text{ of number } n = \sqrt{\mathbf{n}}$$

$$\text{SD of difference between two numbers } n_1 \text{ and } n_2 = (n_1 - n_2)/\sqrt{(n_1 + n_2)}$$

$$\text{SD proportion} = \sqrt{p(1-p)}$$

$$\text{SEM proportion} = \sqrt{p(1-p)/n}$$

We assume that the distribution of proportions of many samples follows a normal distribution (in this case called the **z**-distribution) with 95% confidence intervals between:

$$p \pm 2\sqrt{p\,(1-p)/n}$$

a formula looking very similar to the 95% CI intervals formula for continuous data

$$\text{mean} \pm 2\sqrt{SD^2/n}$$

Differences and sums of the SDs and SEMs of proportions can be calculated similarly to those of continuous data:

$$SEM_{\text{of differences}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\text{with 95\% CI intervals : } \quad p_1 - p_2 \pm 2.\ \text{SEMs}$$

More often than with continuous data, proportions of different samples are assessed for their ratios rather than difference or sum. Calculating the 95% CI intervals of it is not simple. The problem is that the ratios of many samples do not follow a normal distribution, and are extremely skewed. It can never be less than 0 but can get very high. However, the logarithm of the relative risk is approximately symmetrical. Katz's method takes advantage of this symmetry:

$$95\%\ \text{CI of } \log RR = \log RR \pm 2\sqrt{\frac{b/a}{a+b} + \frac{d/c}{c+d}}$$

This equation calculates the CIs of the logarithm of the RR. Take the antilogarithm (10^x) to determine the 95% CIs of the RR.

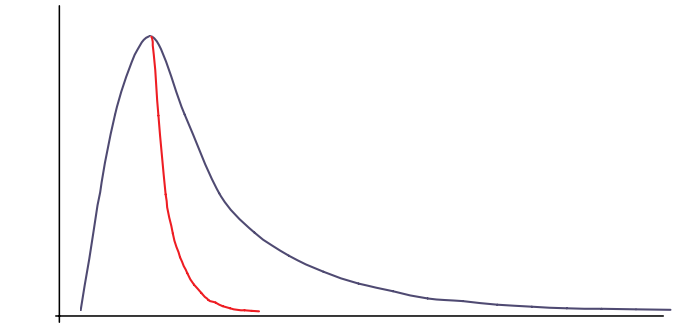


Figure 4. Ratios of proportions unlike continuous data usually do not follow a normal but a skewed distribution (values vary from 0 to ∞). Transformation into the logarithms provides approximately symmetric distributions (thin curve).

Figure 4 shows the distribution of RRs and the distribution of the logarithms of the RRs, and illustrates that the transformation from skewed data into their logarithms is a useful method to obtain an approximately symmetrical distribution, that can be analyzed according to the usual approach of SDs, SEMs and CIs.

5. DIFFERENT TYPES OF DATA: CORRELATION COEFFICIENT

The SD and SEM of paired data includes a term called r as described above. For the calculation of r, otherwise called R, we have to take into account that paired comparisons, e.g., those of two drugs tested in one subject generally have a different variance from those of comparison of two drugs in two different subjects. This is so, because between subjects variability of symptoms is eliminated and because the chance of a subject responding beneficially the first time is more likely to respond beneficially the second time as well. We say there is generally a positive correlation between the responses of one subject to two treatments.

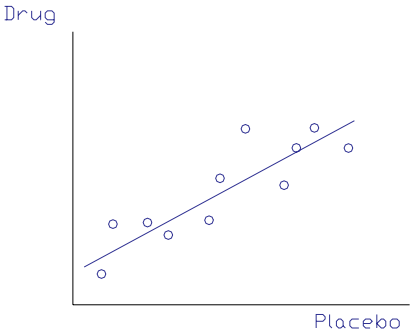


Figure 5. A positive correlation between the response of one subject to two treatments.

Figure 5 gives an example of this phenomenon. X-variables, e.g., blood pressures after the administration of compound 1 or placebo, y-variables blood pressures after the administration of compound 2 or test-treatment.
The SDs and SEMs of the paired sums or differences of the x- and y-variables are relevant to estimate variances in the data and are just as those of continuous data needed before any statistical test can be performed. They can be calculated according to:

$$SD_{\text{paired sum}} = \sqrt{(SD_1^2 + SD_2^2 + 2\,r\,SD_1 \cdot SD_2)}$$
$$SD_{\text{paired differrence}} = \sqrt{(SD_1^2 + SD_2^2 - 2\,r\,SD_1 \cdot SD_2)}$$

where r = correlation coefficient, a term that will be explained soon.

Likewise:

$$SEM_{\text{paired sum}} = \sqrt{(SD_1^2 + SD_2^2 + 2\,r\,SD_1 \cdot SD_2)/n}$$
$$SEM_{\text{paired differrence}} = \sqrt{(SD_1^2 + SD_2^2 - 2\,r\,SD_1 \cdot SD_2)/n}$$

where $n = n_1 = n_2$

and that:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

r is between –1 and +1, and with unpaired data r=0 and the SD and SEM formulas reduce accordingly (as described above). The figure also shows a line, called the

regression line, which presents the best-fit summary of the data, and is the calculated method that minimizes the squares of the distances from the line.

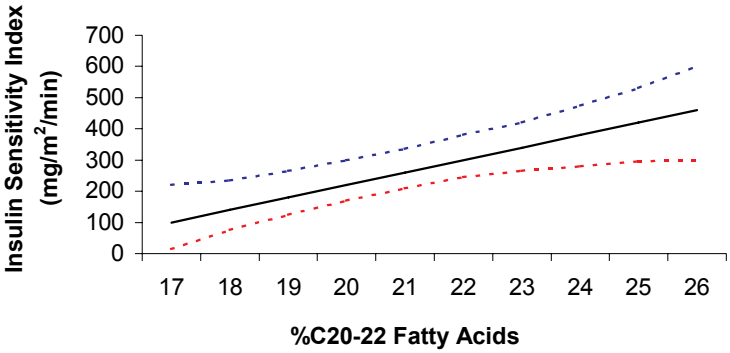


Figure 6. Example of a linear regression line of 2 paired variables (x- and y-values), the regression line provides the best fit line. The dotted curves are 95% CIs that are curved, although we do not allow for a nonlinear relationship between x and y variables.

The 95% CIs of a regression line can be calculated and is drawn as area between the dotted lines in Figure 6. It is remarkable that the borders of the straight regression line are curved although we do not allow for a nonlinear relationship between the x-axis and y-axis variables. More details on regression analysis will be given in chapters 2 and 3.

In the above few lines we described continuous normally distributed or t-distributed data, and rates and their proportions or percentages. We did not yet address data ordered as ranks. This is a special method to transform skewed data into an approximately normal distribution, and is in that sense comparable with logarithmic transformation of relative risks (RRs). In chapter 3 the tests involving this method will be explained.

6. STRATIFICATION ISSUES

When published, a randomized parallel-group drug trial essentially includes a table listing all of the factors, otherwise called baseline characteristics, known possibly to influence outcome. E.g., in case of heart disease these will probably include apart from age and gender, the prevalence in each group of diabetes, hypertension, cholesterol levels, smoking history. If such factors are similar in the two groups, then we can go on to attribute any difference in outcome to the effect of test-treatment over reference-treatment. If not, we have a problem. Attempts are made to retrieve the situation by multiple variables analysis allocating part of the

differences in outcome to the differences in the groups, but there is always an air of uncertainty about the validity of the overall conclusions in such a trial. This issue is discussed and methods are explained in chapter 8. Here we discuss ways to avoid this problem. Ways to do so, are stratification of the analysis and minimization of imbalance between treatment groups, which are both techniques not well-known. Stratification of the analysis means that relatively homogeneous subgroups are analyzed separately. The limitation of this approach is that it can not account for more than two, maybe three, variables, and that, thus, major covariates may be missed. Minimization can manage more factors. The investigators first classify patients according to the factors they would like to see equally presented in the two groups, then randomly assign treatment so that predetermined approximately fixed proportions of patients from each stratum receive each treatment. With this method the group assignment does not rely solely on chance but is designed to reduce any difference in the distribution of unsuspected contributing determinants of outcome so that any treatment difference can now be attributed to the treatment comparison itself. A good example of this method can be found in a study by Kallis et al. [1] The authors stratified in a study of aspirin versus placebo before coronary artery surgery the groups according to age, gender, left ventricular function, and number of coronary arteries affected. Any other prognostic factors other than treatment can be chosen. If the treatments are given in a double-blind fashion, minimization influences the composition of the two groups but does not influence the chance of one group entering in a particular treatment arm rather than the other.

There is an additional argument in favor of stratification/ minimization that counts even if the risk of significant asymmetries in the treatment groups is small. Some prognostic factors have a particularly large effect on the outcome of a trial. Even small and statistically insignificant imbalances in the treatment groups may then bias the results. E.g., in a study of two treatment modalities for pneumonia[2] including 54 patients, 10 patients took prior antibiotic in the treatment group and 5 did in the control group. Even though the difference between 5/27 and 10/27 is not statistically significant, the validity of this trial was being challenged, and the results were eventually not accepted.

7. RANDOMIZED VERSUS HISTORICAL CONTROLS

A randomized clinical trial is frequently used in drug research. However, there is considerable opposition to the use of this design. One major concern is the ethical problem of allowing a random event to determine a patient's treatment. Freirich[3] argued that a comparative trial, which shows major differences between two treatments, is a bad trial because half of the patients have received an inferior treatment. On the other hand, in a prospective trial randomly assigning treatments avoids many potential biases. Of more concern is the trial in which a new treatment is compared to an old treatment when there is information about the efficacy of the old treatment through historical data. In this situation the use of historical data for comparison with data from the new treatment will shorten the length of the study because all patients can be assigned to the new treatment. The current availability

of multivariable statistical procedures which can adjust the comparison of two treatments for differing presence of other prognostic factors in the two treatment arms, has made the use of historical controls more appealing. This has made randomization less necessary as a mechanism for ensuring comparability of the treatment arms. The weak point in this approach is the absolute faith one has to place in the multivariable model. In addition, some confounding variables e.g., time effects, simply can not be adjusted, and remain unknown. Despite the ethical argument in favor of historical controls we must therefore emphasize the potentially misleading aspects of trials using historical controls.

8. FACTORIAL DESIGNS

The majority of drug trials are designed to answer a single question. However, in practice many diseases require a combination of more than one treatment modalities. E.g., beta-blockers are effective for stable angina pectoris but beta-blockers plus calcium channel blockers or beta-blockers plus calcium channel blockers plus nitrates are better (Table 1). Not addressing more than one treatment modality in a trial is an unnecessary restriction on the design of the trial because the assessment of two or more modalities in on a trial pose no major mathematical problems.

Table 1. The factorial design for angina pectoris patients treated with calcium channel blockers with or without beta-blockers

	Calcium channel blocker	no calcium channel blocker
Beta-blocker	regimen I	regimen II
No beta-blocker	regimen III	regimen I

We will not describe the analytical details of such a design but researchers should not be reluctant to consider designs of such types. This is particularly so, when the recruitment of large samples causes difficulties.

9. CONCLUSIONS

What you should know after reading this chapter:
1. Scientific rules governing controlled clinical trials include prior hypotheses, valid designs, strict description of the methods, uniform data analysis.
2. Efficacy data and safety data often involve respectively continuous and proportional data.
3. How to calculate standard deviations and standard errors of the data.

4. You should have a notion of negative/positive correlation in paired comparisons, and of the meaning of the so-called correlation coefficient.
5. Mean ± standard deviation summarizes the data, mean ± standard error summarizes the means of many trials similar to our trial.
6. You should know the meaning of historical controls and factorial designs.

10. REFERENCES

1. Kallis F et al. Aspirin versus placebo before coronary artery surgery. Eur J Cardiovasc Surg 1994; 8: 404-10.
2. Graham WG, Bradley DA. Efficacy of chest physiotherapy and intermittent positive-pressure breathing in the resolution of pneumonia. N Engl J Med 1978; 299: 624-7.
3. Freirich F. Ethical problem of allowing a random event to determine a patient's treatment. In: Controversies in clinical trials. Saunders, Philadelphia, 1983, p 5.

CHAPTER 2

THE ANALYSIS OF EFFICACY DATA

1. OVERVIEW

Typical efficacy endpoints have their associated statistical techniques. For example, values of continuous measurements (e.g., blood pressures) require the following statistical techniques:

(a) if measurements are normally distributed: t-tests and associated confidence intervals to compare two mean values; analysis of variance (ANOVA) to compare three or more,
(b) if measurements have a non-normal distribution: Wilcoxon or Mann-Whitney tests with confidence intervals for medians.

Comparing proportions of responders or proportions of survivors or patients with no events involves binomial rather than normal distributions and requires a completely different approach. It requires a chi-square test, or a more complex technique otherwise closely related to the simple chi-square test, e.g., Mantel Haenszl summary chi-square test, logrank test, Cox proportional hazard test etc. Although in clinical trials, particularly phase III-IV trials, proportions of responders and proportion of survivors is increasingly an efficacy endpoint, in many other trials proportions are used mainly for the purpose of assessing safety endpoints, while continuous measurements are used for assessing the main endpoints, mostly efficacy endpoints. We will, therefore, focus on statistically testing continuous measurements in this chapter and will deal with different aspects of statistically testing proportions in the next chapter.
Statistical tests all have in common that they try to estimate the probability that a difference in the data is true rather than due to chance. Usually statistical tests make use of a so-called **test statistic:**

Chi-square	for the chi-square test
t	for the t-test
Q	for nonparametric comparisons
Q^1	for nonparametric comparisons
q	for Newman-Keuls test
q^1	for Dunnett test
F	for analysis of variance
Rs	for Spearman rank correlation test.

These test statistics can adopt different sizes. In the appendix of this book we present tables for t-, chi-square- and F- , Mann-Whitney-, and Wilcoxon-rank-sum-

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 17–43.

tests, but additional tables are published in most textbooks of statistics (see References). Such tables show us the larger the size of the test statistic, the more likely it is that the null-hypothesis of no difference from zero or no difference between two samples is untrue, and that there is thus a true difference or true effect in the data. Most tests also have in common that they are better sensitive or powerful to demonstrate such a true difference as the samples tested are large. So, the test statistic in most tables is adjusted for sample sizes. We say that the sample size determines the degrees of freedom, a term closely related to the sample size.

2. THE PRINCIPLE OF TESTING STATISTICAL SIGNIFICANCE

The human brain excels in making hypotheses but hypotheses may be untrue. When you were a child you thought that only girls could become a doctor because your family doctor was a female. Later on, this hypothesis proved to be untrue. Hypotheses must be assessed with hard data. Statistical analyses of hard data starts with assumptions:

1. our study is representative for the entire population (if we repeat the trial, difference will be negligible.
2. All similar trials will have the same standard deviation (SD) or standard error of the mean (SEM).

Because biological processes are *full* of variations, statistics will give no certainties only chances. What chances? Chances that hypotheses are true / untrue. What hypotheses?: e.g.:

(1) our mean effect is not different from a 0 effect,
(2) it is really different from a 0 effect,
(3) it is worse than a 0 effect.

Statistics is about estimating such chances / testing such hypotheses. Please note that trials often calculate differences between a test treatment and a control treatment and, subsequently, test whether this difference is larger than 0. A simple way to reduce a study of two groups of data and, thus, two means to a single mean and single distribution of data, is to take the difference between the two and compare it with 0.

In the past chapter we explained that the data of a trial can be described in the form of a normal distribution graph with SEMs on the x-axis, and that this method is adequate to test various statistical hypotheses. We will now focus on a very important hypothesis, the null-hypothesis. What it literally means is: no difference from a 0 effect: the mean value of our sample is not different from the value 0. We will try and make a graph of this null-hypothesis.

What does it look like in graph? H1 in Figure 1 is a graph based on the data of our trial with SEMs distant from mean on the x-axis (z-axis). H0 is the same graph with a mean value of 0 (mean ± SEM= 0 ± 1). Now, we will make a giant leap

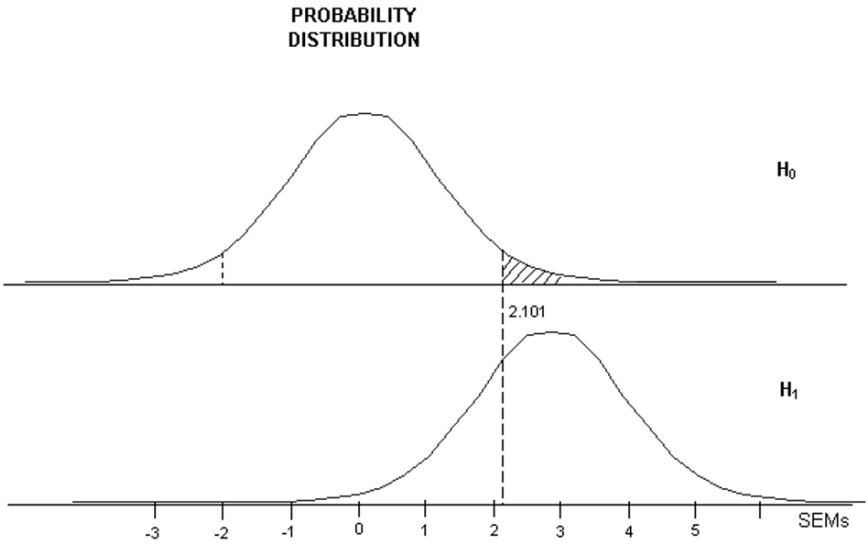


Figure 1. Null-hypothesis (H_0) and alternative hypothesis H_1 of an example of experimental data with sample size (n) = 20 and mean = 2.9 SEMs, and a t-distributed frequency distribution.

from our data to the entire population, and we can do so, because our data are representative for the entire population. H1 is also the summary of the means of many trials similar to ours (if we repeat, differences will be small, and summary will look alike). H0 is also the summary of the means of many trials similar to ours but with an overall effect of 0. Now our mean effect is not 0 but 2.9. Yet it could be an outlier of many studies with an overall effect of 0. So, we should think from now on of H0 as the distribution of the means of many trials with overall effect of 0. If H0 is true, then the mean of our study is part of H0. We can not prove anything, but we can calculate the chance/probability of this possibility.

A mean value of 2.9 is far distant from 0. Suppose it belongs to H0. Only 5% of the H0 trials have their means >2.1 SEMs distant from 0, because the area under the curve (AUC) >2.1 distant from 0 is only 5% of total AUC. Thus, the chance that our mean belongs to H0 is <5%. This is a small chance, and we reject this chance and conclude there is <5% chance to find this result. We, thus, reject the H0 of no difference from 0 at P<0.05. The AUC right from 2.101 (and left from -2.101 as will be soon explained) is called alpha= area of rejection of H0. Our result of 2.9 is far from 2.101. The probability of finding such a result may be a lot smaller than 5%. Table 1 shows the t-table that can tell us exactly how small this chance truly is.

Table 1. t-table

	Two-tailed P-value (df = degree of freedom)			
df	**0.1**	**0.05**	**0.01**	**0.002**
1	6.314	12.706	63.657	318.31
2	2.920	4.303	9.925	22.326
3	2.353	3.182	5.841	10.213
4	2.132	2.776	4.604	7.173
5	2.015	2.571	4.032	5.893
6	1.943	2.447	3.707	5.208
7	1.895	2.365	3.499	4.785
8	1.860	2.306	3.355	4.501
9	1.833	2.262	3.250	4.297
10	1.812	2.228	3.169	4.144
11	1.796	2.201	3.106	4.025
12	1.782	2.179	3.055	3.930
13	1.771	2.160	3.012	3.852
14	1.761	2.145	2.977	3.787
15	1.753	2.131	2.947	3.733
16	1.746	2.120	2.921	3.686
17	1.740	2.110	2.898	3.646
18	1.734	2.101	2.878	3.610
19	1.729	2.093	2.861	3.579
20	1.725	2.086	2.845	3.552
21	1.721	2.080	2.831	3.527
22	1.717	2.074	2.819	3.505
23	1.714	2.069	2.807	3.485
24	1.711	2.064	2.797	3.467
25	1.708	2.060	2.787	3.450
26	1.706	2.056	2.779	3.435
27	1.701	2.052	2.771	3.421
28	1.701	2.048	2.763	3.408
29	1.699	2.045	2.756	3.396
30	1.697	2.042	2.750	3.385
40	1.684	2.021	2.704	3.307
60	1.671	2.000	2.660	3.232
120	1.658	1.950	2.617	3.160
∞	1.645	1.960	2.576	3.090

The 4 right-hand columns are trial results expressed in SEM-units distant from 0 **(=also t-values)**. The upper row gives the AUC-values right from trial results. The left-hand column presents adjustment for numbers of patients (degrees of freedom (dfs), in our example two samples of 10 gives 20-2:= 18 dfs).
AUC right from 2.9 means→ right from 2.878 means→this AUC<0.01. And so we conclude that our probability not <0.05 but even<0.01.Note: the t-distribution is just an adjustment of the normal distribution, but a bit wider for small samples. With large samples it is identical to the normal distribution. For proportional data always the normal distribution is applied.

Note: Unlike the t-table in the APPENDIX, the above t-table gives two-tailed = two-sided AUC-values. This means that the left and right end of the frequency distribution are tested simultaneously. A result >2.101 here means both >2.101 and <-2.101. If a result of + 2.101 was tested one sided, the p-value would be 0.025 instead of 0.05 (see t-table APPENDIX).

3. THE T-VALUE = STANDARDIZED MEAN RESULT OF STUDY

The t-table expresses the mean result of a study in SEM-units. Why does it make sense to express mean results in SEM-units? Consider a cholesterol reducing compound, which reduces plasma cholesterol by 1.7 mmol/l ± 0.4 mmol/l (mean ± SEM). Is this reduction statistically significant? Unfortunately, there are no statistical tables for plasma cholesterol values. Neither are there tables for blood pressures, body weights, hemoglobin levels etc. The trick is to standardize your result.

$$\text{Mean} \pm \text{SEM} = \frac{\text{Mean}}{\text{SEM}} \pm \frac{\text{SEM}}{\text{SEM}} = \text{t - value} \pm 1$$

This gives us our test result in SEM-units with an SEM of 1. Suddenly, it becomes possible to analyze every study by using one and the same table, the famous t-table. How do we know that our data follow a normal or t frequency distribution? We have goodness of fit tests (chapter 24).
How was the t-table made? It was made in an era without pocket calculators, and it was hard work. Try and calculate in three digits the square root of the number 5. The result is between 2 and 3. The final digits are found by a technique called “tightening the data”. The result is larger than 2.1, smaller than 2.9. Also larger than 2.2, smaller than 2.8, etc. It will take more than a few minutes to find out the closest estimate of $\sqrt{5}$ in three digits. This example highlights the hard work done by the U.S. Government’s Work Project Administration by hundreds of women during the economic depression in the 1930s.

4. UNPAIRED T-TEST

So far, we assessed a single mean versus 0, now we will assess two means versus each other. For example, a parallel-group study of two groups tests the effect of two beta-blockers on cardiac output.

	Mean ± SD			$SEM^2 = SD^2 / n$
group 1 (n =10)	5.9	± 2.4	liter / min	5.76 / 10
group 2 (n =10)	4.5	± 1.7	liter / min	2.89 / 10

Calculate: $mean_1 - mean_2$ = mean difference =1.4
Then calculate pooled $SEM = \sqrt{SEM_1^2 + SEM_2^2} = 0.930$

Note: for SEM of difference: take the square root of the sums of squares of separate SEMs and so reduce analysis of two means and two SEMS to one mean and one SEM. The significance of difference between two unpaired samples of continuous data is assessed by the formula:

$$mean_1 - mean_2 \pm \sqrt{SEM_1^2 + SEM_2^2} = \text{mean difference} \pm \text{pooled SEM}$$

This formula presents again a t-distribution with a new mean and a new SEM, i.e., the mean difference and the pooled SEM. The wider this new mean is distant from zero and the smaller its SEM is, the more likely we are able to demonstrate a true effect or true difference from no effect. The size of the test statistic is calculated as follows.

$$\text{The size of } t = \frac{\text{mean difference}}{\text{pooled SEM}} = 1.4 / 0.930 = 1.505$$

With n=20, and two groups we have 20-2 = 18 degrees of freedom. The t-table shows that a t-value of 1.505 provides a chance of >5% that the null hypothesis of no effect can be rejected. The null-hypothesis cannot be rejected.

Note: If the standard deviations are very different in size, e.g., if one is twice the other, then a more adequate calculation of the pooled standard error is as follows.

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

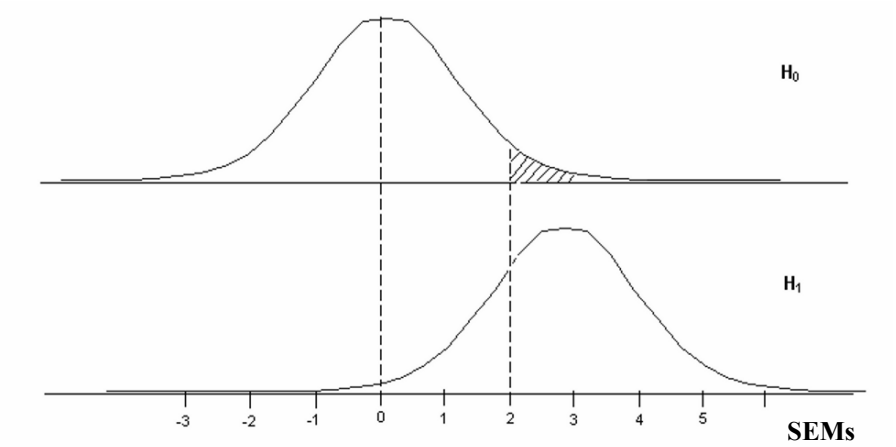


Figure 2. Two t- distributions with n=20: lower curve H1 or actual SEM-distribution of the data, upper curve H0 or null hypothesis of the study.

The lower graph of Figure 2 is the probability distribution of this t-distribution. H0 (the upper graph) is an identical distribution with mean = 0 instead of mean = $mean_1$-$mean_2$ and with SEM identical to the SEM of H1, and is taken as the null-hypothesis in this particular approach. With n=20 (18 dfs) we can accept that 95% of all t-distributions with no significant treatment difference from zero must have their means between –2.101 and +2.101 SEMs distant from zero. The chance of finding a mean value of 2.101 SEMs or more distant from 0 is 5% or less (we say α=0.05, where α is the chance of erroneously rejecting the null hypothesis of no effect). This means that we can reject the null-hypothesis of no difference at a probability (P) =0.05. We have 5% chance of coming to this result, if there were no difference between the two samples. We, therefore, conclude that there is a true difference between the effects on cardiac output of the two compounds.

Also the F- and chi-square test reject, similarly to the t-test, reject the null-hypothesis of no treatment effect if the value of the test statistic is larger than would occur in 95% of the cases if the treatment had no effect. At this point we should emphasize that when the test statistic is not big enough to reject the null-hypothesis of no treatment effect, investigators often report no statistically significant difference and discuss their results in terms of documented proof that the treatment had no effect. All they really did, was, fail to demonstrate that it did have an effect. The distinction between positively demonstrating that a treatment had no effect and failing to demonstrate that it does have an effect, is subtle but very important, especially with respect to the small numbers of subjects usually enrolled in a trial. A study of treatments that involves only a few subjects and then fails to reject the null-hypothesis of no treatment effect, may arrive at that conclusion because the statistical procedure lacked power to detect the effect

because of a too small sample size, even though the treatment did have an effect. We will address this problem in more detail in chapter 6.

5. NULL-HYPOTHESIS TESTING OF 3 OR MORE UNPAIRED SAMPLES

If more than two samples are compared, things soon get really complicated, and the unpaired t-test can no longer be applied. Usually, statistical software, e.g., SAS or SPSS Statistical Software, will be used to produce F- or P-values, but the Table 2 gives a brief summary of the principles of multiple groups analysis of variance (ANOVA) applied for this purpose. With ANOVA the outcome variable (Hb, hemoglobin-level in the example) is often called the dependent variable, while the groups-variable is called the independent factor (SPSS[1]: Compare means; one-way ANOVA). If additional groups-variables are in the data (gender, age classes, comorbidities), then SPSS requires using the General Linear Model (univariate).

Table 2. Multiple groups ANOVA

Unpaired ANOVA 3 groups

Total variation

| |

Between group variation within group variation

In ANOVA:

Variations are expressed as sums of squares (SS) and can be added up to obtain total variation. Assess whether between-group variation is large compared to within-group variation.

Group	n patients	mean	SD
1	-	-	-
2	-	-	-
3	-	-	-

Grand mean = (mean 1 + 2 +3)/3

$SS_{\text{between groups}} = n\,(\text{mean}_1 - \text{grand mean})^2 + n\,(\text{mean}_2 - \text{grand mean})^2 + \ldots$

$SS_{\text{within groups}} = (n-1)(SD_1^2) + (n-1)\,SD_2^2 + \ldots..$

$$F = \frac{\text{SS between groups /dfs}}{\text{SS within groups / dfs}} = MS_{\text{between}} / MS_{\text{within}}$$

F-table gives P-value

Effect of 3 compounds on Hb

Group	n patients	mean	SD
1	16	8.7125	0.8445
2	16	10.6300	1.2841
3	16	12.3000	0.9419

Grand mean = (mean 1 + 2 +3)/3 = 10.4926

$SS_{between\ groups}$ = 16 (8.7125-10.4926) ^{2}+ 16(10.6300 – 10.4926)^2 +….
$SS_{within\ groups}$ = 15 x 0.8445^2 + 15 x 1.2841^2 +……..
F = 49.9 and so P <0.001

Note: In case 2 groups: ANOVA= unpaired T-test ($F=T^2$). Dfs means degrees of freedom, and equals 3n - 3 for SS_{within} , and 3-1=2 for $SS_{between}$.

6. THREE METHODS TO TEST STATISTICALLY A PAIRED SAMPLE

Table 3 gives an example of a placebo-controlled clinical trial to test efficacy of a sleeping drug.

Table 3. Example of a placebo-controlled clinical trial to test efficacy of a sleeping drug

	hours of sleep				
patient	drug	placebo	difference	mean	SS
1	6.1	5.2	0.9	5.7	0.53
2	7.0	7.9	–0.9	7.5	
3	8.2	3.9	4.3		
4	7.6	4.7	2.9		
5	6.5	5.3	1.2		
6	7.8	5.4	3.0		
7	6.9	4.2	2.7		
8	6.7	6.1	0.6		
9	7.4	3.8	3.6		
10	5.8	6.3	–0.5		
Mean	7.06	5.28	1.78		
SD	0.76	1.26	1.77		
grand mean	6.17				

First method
First method is simply calculating the SD of the mean difference d by looking at the column of differences (d-values) and using the standard formula for variance between data

$$\text{SD paired differences} = \sqrt{\frac{\sum (d-\bar{d})^2}{n-1}} = 1.79$$

Next we find SEM of the mean difference by taking SD/$\sqrt{n}$ =0.56

Mean difference ± SEM= 1.78 ± 0.56

Similarly to the above unpaired t-test we now can test the null hypothesis of no difference by calculating

$$t = \frac{\text{Mean difference}}{\text{SEM}} = 1.78 / 0.56 = 3.18 \text{ with a sample of 10 (degrees of freedom = 10 -1)}$$

The t-table shows that P<0.02. We have <2% chance to find this result if there were no difference, and accept that this is sufficient to assume that there is a true difference.

Second method

Instead of taking the column of differences we can take the other two columns and use the formula as described in chapter 1 for calculating the SD of the paired differences=$SD_{\text{paired differrence}}$

$$= \sqrt{(SD_1^2 + SD_2^2 - 2r \cdot SD_1 \cdot SD_2)}$$

$$= \sqrt{(0.76_1^2 + 1.26_2^2 - 2r \cdot 0.76_1 \cdot 1.26_2)}$$

As r can be calculated to be +0.26, we can now conclude that

$SD_{\text{paired differrence}}$ = 1.79

The remainder of the calculations is as above.

Third method

The third method is the F test using analysis of variance (ANOVA). We have to calculate SS (sum of squares) e.g., for subject 1:

$SS_{\text{within subject 1}}$ = $(6.1-5.7)^2$+ $(5.2-5.7)^2$ +…..= 0.53 (table 3)

grand mean (7.06+5.28)/ 2= 6.17 (table 3)

$SS_{\text{within subject}}$ = $SS_{\text{within subject 1}}$+ $SS_{\text{within subject 2}}$+$SS_{\text{within subject 3}}$+….

$SS_{\text{treatment}}$= $(7.06-6.17)^2$+$(5.28-6.17)^2$ (table 3)

SS_{residual}= $SS_{\text{within subject}}$ − $SS_{\text{treatment}}$

Table 4. ANOVA table of these data

Source of variation	Sum of Squares (SS)	degrees of freedom (dfs)	mean square MS=SS/dfs)	F = MS treatment / MS residual
between subjects		2 (m)		
within subjects		10 (n x (m-1))		
treatments		1 (m-1)		F= 10.11, p<0.02

residual	9	(n-1)
total	22	

The ANOVA table (Table 4) shows the procedure. Note m is number of treatments, n is number of patients. The ANOVA is valid not only for two repeated measures but also for m repeated measures. For 2 repeated measures it is actually equal to the paired t-test (= first method). The results of the analysis of the two tests are similar, with F being equal to t^2 .
Similarly, for unpaired samples, with two samples the one way ANOVA already briefly mentioned in chapter 1 is equal to the unpaired t-test, but one-way ANOVA can also be used for *m* unpaired samples.

The above data can also be presented in the form of a linear regression graph.

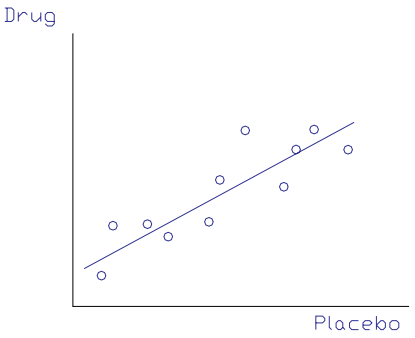


Figure 3. Paired data laid out in the form of linear regression.

Paired data can also be laid out in the form of linear regression

$$y=a+bx \qquad \text{(effect drug)} = a + b\ \text{(effect placebo)}$$

which can be assessed in the form of ANOVA:

$$F = r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{(\sum (x - \bar{x})(y - \bar{y}))^2}{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2} = \frac{SP^2\ x \cdot y\text{ - values}}{SS\ x\text{ - values} \cdot SS\ y\text{ - values}}$$

SS regression= SP^2 x . y -values / SS x -values
SS total = SS y
SS regression / SS total = r^2
SP indicates sum of products.

Table 5. ANOVA table for the linear regression between paired samples

Source of variation	Sum of Squares (SS)	degrees of freedom (dfs)	mean square MS=SS/dfs	F = MS regression / MS total
regression between samples	1.017	1	1.017	0.61, P>0.05
residual	14.027	8	1.753	
total	15.044	9	1.672	

The ANOVA table (Table 5) gives an alternative interpretation of the correlation coefficient; the square of the correlation coefficient, r, equals the regression sum of squares divided by the total sum of squares (0.26^2 = 0.0676= 1.017/15.044) and, thus, is the proportion of the total variation that has been explained by the regression. We can say that the variances in the drug data are only for 6.76% determined by the variances in the placebo data, and that they are for 93.24% independent of the placebo data. With strong positive correlations, e.g., close to +1 the formula for SD and thus SEM reduces to a very small size (because [SD_1^2 + SD_2^2 – 2 r SD_1. SD_2] will be close to zero), and the paired t-test produces huge sizes of t and thus huge sensitivity of testing. The above approach cannot be used for estimating significance of differences between two paired samples. And the method in the presented form is not very relevant. It starts, however, to be relevant, if we are interested in the dependency of a particular outcome variable upon several factors. E.g., the effect of a drug is better than placebo but this effect still gets better with increased age. This concept can be represented by a multiple regression equation

$$y= a+b_1x_1+b_2x_2$$

which in this example is

drug response= a + b_1 . (placebo response) + b_2 . (age)

Although it is no longer easy to visualize the regression, the principles involved are the same as with linear regression. In the Chapters 13 and 14 this subject will be dealt with more explicitly.

7. NULL-HYPOTHESIS TESTING OF 3 OR MORE PAIRED SAMPLES

If more than two paired samples are compared, things soon get really complicated, and the paired t-test can no longer be applied. Usually, statistical software (SAS, SPSS)[1] will be used to produce F- and P-values, but the Table 6 gives a brief summary of the principles of ANOVA for multiple paired observations, used for this purpose. A more in-depth treatment of repeated measures methods will be given in Chapter 38.

Table 6. Repeated measurements ANOVA

Paired ANOVA 3 treatments in single group

Total variation
| |
Between subject variation Within-subject variation
| |
Between treatment variation Residual variation

Variations expressed as sums of squares (SS) and can be added up
Assess whether between treatment variation is large compared to residual variation.

Subject	treatment 1	treatment 2	treatment 3	SD^2
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
Treatment mean	-	-	-	

Grand mean = (treatment mean 1 + 2 + 3)/ 3= …..

$SS_{within\ subject} = SD_1^2 + SD_2^2 + SD_3^2 + ..$
$SS_{treatment}$ = (treatment mean 1 – grand mean)^2 + (treatment mean 2- grand mean)^2 +…..
$SS_{residual} = SS_{within\ subject} - SS_{treatment}$

$$F = \frac{SS_{treatment}/dfs}{SS_{residual}/dfs}$$

F table gives P-value.

Effect of 3 treatments on vascular resistance (blood pressure / cardiac output).

Person	treatment 1	treatment 2	treatment 3	SD^2
1	22.2	5.4	10.6	147.95
2	17.0	6.3	6.2	77.05
3	14.1	8.5	9.3	18.35

4	17.0	10.7	12.3	21.4
Treatment mean	17.58	7.73	9.60	

Grand mean = 11.63

$SS_{within\ subj} = 147.95 + 77.05 + \ldots$
$SS_{treatment} = (17.58 - 11.63)^2 + (7.73 - 11.63)^2 + \ldots$
$SS_{residual} = SS_{within\ subject} - SS_{treatment}$

F= 18.2 and so P<0.025

Note: in case of 2 treatments: repeated measurements-ANOVA produces the same result as the paired t-test ($F = t^2$), dfs= degrees of freedom equals 3-1=2 for $SS_{treatment}$, and 4-1=3 for $SS_{residual}$.

8. NULL-HYPOTHESIS TESTING WITH COMPLEX DATA

ANOVA is briefly addressed in the above sections 6 and 7. It is a powerful method for the analysis of complex data, and will be addressed again in many of the following chapters of this book. ANOVA compares mean values of multiple cells, and can be classified in several manners: (1) one-way or two-way (Table 7, left example gives one-way ANOVA with 3 cells, right example two-way ANOVA with 6 cells), (2) unpaired or paired data, if the cells contain either non-repeated or partly repeated data (otherwise called repeated measures ANOVA), (3) data with or without replication, if the cells contain either multiple data or a single datum, (4) balanced or unbalanced, if the cells contains equal or differing numbers of data.

Table 7. ANOVA compares multiple cells with means, and can be classified in several ways

(1) One-way		Two-way			
	mean blood pressure		mean results of treatments 1-3		
group 1	….. (SD…)		1	2	3
group 2	……	males	….	…..	…..
group 3	…..	females	…..	……	…..
(2) unpaired data		unpaired data / paired data			
(3) with replication		with replication / without replication			
(4) balanced / unbalanced		balanced / unbalanced			

Sometimes samples consist of data that are partly repeated and partly non-repeated. E.g., 10 patients measured 10 times produces a sample of n=100. It is not appropriate to include this sample in an ANOVA-model as either entirely repeated or non-repeated. It may be practical, then, to use the means per patient as a summary measure without accounting its standard deviation, and perform simple tests using the summary measures per patient only. Generally, the simpler the statistical test the more statistical power.

9. PAIRED DATA WITH A NEGATIVE CORRELATION

Many crossover and parallel-group studies include an element of self-controlling. E.g., observations before, during, and after treatment are frequently used as the main control on experimental variation. Such repeated measures will generally have a positive correlation: those who respond well during the first observation are more likely to do so in the second. This is, however, not necessarily so. When drugs of completely different classes are compared, patients may fall apart into different populations: those who respond better to one and those who respond better to the other drug. For example, patients with angina pectoris, hypertension, arrhythmias, chronic obstructive pulmonary disease, unresponsive to one class of drugs, may respond very well to a different class of drugs. This situation gives rise to a negative correlation in a paired comparison. Other examples of negative correlations between paired observations include the following. A negative correlation between subsequent observations in one subject may occur because fast-responders are more likely to stop responding earlier. A negative correlation may exist in the patient characteristics of a trial, e.g., between age and vital lung capacity, and in outcome variables of a trial, e.g., between severity of heart attack and ejection fraction. Negative correlations in a paired comparison reduce the sensitivity not only of studies testing differences but also of studies testing equivalences (Chapter 4).

I. Studies testing significance of differences

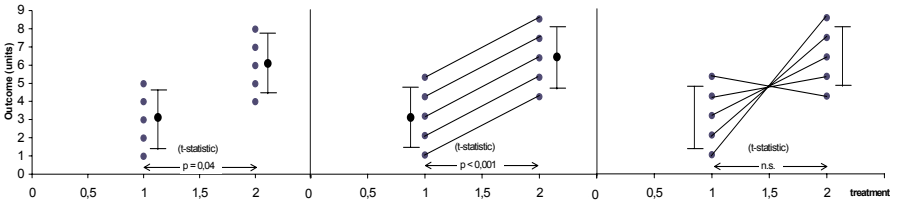


Figure 4. Hypothesized examples of three studies: left graph parallel-group study of 10 patients, middle and right graphs self-controlled studies of 5 patients each tested twice.

Figure 4 gives a hypothesized example of three studies: the left graph shows a parallel-group study of 10 patients, the middle and right graph show self-controlled studies of 5 patients each tested twice. T-statistics is employed according to the formula

$$t = \frac{\bar{d}}{SE}$$

Where $\bar{d}$ is the mean difference between the two sets of data (6-3=3) and the standard error (SE) of this difference is calculated for the left graph data according to

$$\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} = 0.99$$

$SD_1$and SD_2 are standard deviations and n_1 and n_2 are numbers of observations in each of the groups. We assume that $n_1 = n_2 = n$.

$$t = 3 / 0.99 = 3.0$$

With 10 observations we can reject the null-hypothesis at p = 0.04.
With a positively paired comparison (middle graph) we have even more sensitivity. SE is calculated slightly different

$$SE = \frac{\sqrt{\sum (d - \bar{d})^2 / (n-1)}}{\sqrt{n}} = 0$$

where d is the observed change in each individual and $\bar{d}$ is its mean.

$$t = \bar{d} / SE = 3/0 = \infty$$

with n=5 we can reject the null-hypothesis at p<0.001.

The right graph gives the negative correlation situation. SE calculated similarly to the middle graph data is 1.58, which means that

$$t = 3 / 1.58 = 1.89$$

The null-hypothesis of no difference cannot be rejected. Differences are not significant (n.s.).
When more than 2 treatments are given to one sample of patients t-statistics is not appropriate and should be replaced by analysis of variance.

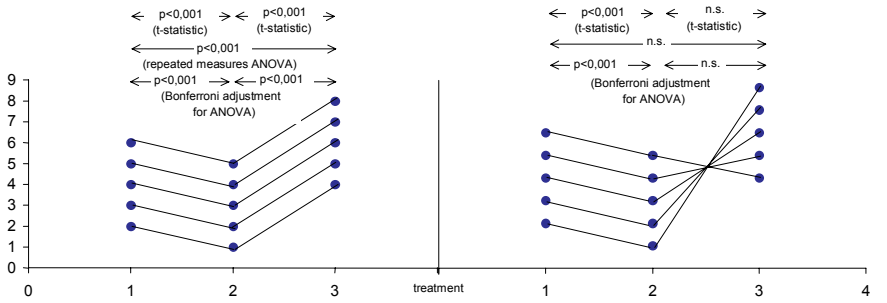


Figure 5. Hypothesized example of two studies where 5 patients are tested three times. Due to negative correlation between treatment 2 and 3 in the right study, the statistical significance test is negative unlike the left graph study, despite the identical mean results.

Figure 5 gives a hypothesized example of two studies where 5 patients are tested three times. In the left graph the correlation between treatment responses is positive, whereas in the right graph the correlation between treatment no.3 and no.2 is strong negative rather than positive. For the left graph data repeated measures ANOVA is performed.
The sum of squares (SS) of the different treatments is calculated according to

Patient	treatment 1	treatment 2	treatment 3	Mean	SD^2
1	6	5	8	6.3	4.67
2	5	4	7	5.3	4.67
3	4	3	6	4.3	4.67
4	3	2	5	3.3	4.67
5	2	1	4	2.3	4.67
Treatment mean	4	3	6		

Grand mean 4.3

$SS_{\text{within subjects}} = 4.67 + 4.67 + \ldots = 23.3$

$SS_{\text{treatments}} = 5\,[(4-4.3)^2+(3-4.3)^2+(6-4.3)^2] = 23.35$

$SS_{\text{residual}} = SS_{\text{within subjects}} - SS_{\text{treatments}} = 0$

Table 8. ANOVA table of the data

Source of variation	SS	dfs	MS
Within subjects	23.35	10	
Treatments	23.35	2	11.68
Residual	0	8	0

$$F = \frac{MS_{treatments}}{MS_{residual}} = \infty \qquad p<0.001$$

This analysis permits concluding that at least one of the treatments produces a change. To isolate which one, we need to use a multiple-comparisons procedure, e.g., the modified Bonferroni t-test for ANOVA where

"$SE^2 = \Sigma(d - \overline{d})^2 / (n-1)$" is replaced with "$MS_{residual}$" (Table 8). So, to compare, e.g., treatment no. 2 with treatment no. 3

$$t = \frac{6 - 3}{\sqrt{(MS_{residual})/n}} = \infty \qquad p<0.001$$

Of the right graph from (Figure 5) a similar analysis is performed.

Patients	treatment 1	treatment 2	treatment 3	Mean	SD^2
1	6	5	4	5.0	1.0
2	5	4	5	4.7	0.67
3	4	3	6	4.3	4.67
4	3	2	7	4.0	14.0
5	2	1	8	3.7	28.49
Treatment mean	4	3	6		

Grand mean 4.3

$SS_{within\ subjects} = 1.0 + 0.67 + 4.67 + \ldots = 48.83$
$SS_{treatments} = 5\,[(4-4.3)^2+(3-4.3)^2+(6-4.3)^2] = 23.35$
$SS_{residual} = SS_{within\ subjects} - SS_{treatments} = 48.83 - 23.35 = 24.48$

Table 9. ANOVA table of the data

Source of variation	SS	DF	MS
Within subjects	48.83	10	
Treatments	23.35	2	11.7
Residual	24.48	8	3.1

$$F = \frac{MS_{treatments}}{MS_{residual}} = 3.77 \qquad p=0.20$$

This analysis does not permit concluding that one of the treatments produces a change (Table 9). The Bonferroni adjustment of treatments no. 2 and no. 3 of course, does not either (p=0.24 and p=0.34).

In conclusion, with negative correlations between treatment responses statistical methods including paired t-statistics, repeated measures ANOVA, and Bonferroni adjustments for ANOVA lack sensitivity to demonstrate significant treatment effects. The question why this is so, is not difficult to recognize. With t-statistics and a negative correlation between-patient-variation is almost doubled by taking paired differences. With ANOVA things are similar.

$SS_{within\ subjects}$ are twice the size of the positive correlation situation while $SS_{treatments}$ are not different. It follows that the positive correlation situation provides a lot more sensitivity to test than the negative correlation situation.

II. Studies testing equivalence

In an equivalence trial the conventional significance test has little relevance: failure to detect a difference does not imply equivalence, and a difference, which is detected may not have any clinical relevance and, thus, may not correspond to clinically relevant equivalence. In such trials the range of equivalence is usually predefined as an interval from –D to +D distant from a difference of 0. D is often set equal to a difference of undisputed clinical importance, and hence may be above the minimum of clinical interest by a factor two or three. The bioequivalence study design essentially tests both equivalence and superiority / inferiority. Let us assume that in an equivalence trial of vasodilators for Raynaud's phenomenon 10 patients are treated with vasodilator 1 for one week and for a separate period of one week with vasodilator 2. The data below show the numbers of Raynaud attacks per week (Table 10).

Table 10. Correlation levels and their influence on sensitivity of statistical tests

ρ = -1			ρ = 0			ρ =+1		
vasodilator			vasodilator			vasodilator		
one	two	paired differences	one	two	paired differences	one	two	paired differences
45	10	35	45	40	5	10	10	0
40	15	25	40	35	5	20	15	5
40	15	25	40	35	5	25	15	10
35	20	15	35	30	5	25	20	5
30	25	5	30	25	5	30	25	5
30	25	5	30	10	20	30	25	5
25	30	-5	25	15	10	35	30	5
25	35	-10	25	15	10	40	35	5
20	35	-15	20	20	0	40	35	5
10	40	-30	10	25	-15	40	40	5
means								
30	25	5	30	25	5	30	25	5
SEMs								
3.16	3.16	6.46	3.16	3.16	2.78	3.16	3.16	0.76
t-values								
0.8			1.8			6.3		
95% CIs								
± 14.5			±6.3			± 1.7		

SEM=standard error of the mean;
t means level of t according to t-test for paired differences;
CI means confidence interval calculated according to critical t value of t-distribution for 10-1
pairs = 9 degrees of freedom (critical t =2.26, 95% CI= 2.26 x SEM);
ρ= correlation coefficient (the Greek letter is often used instead of r if we mean total population instead of our sample).

Although samples have identical means and SEMs (25 ± 3.16 x-axis, 30 ± 3.16 y-axis) their correlation coefficients range from –1 to +1. The null hypothesis of no equivalence is rejected when the 95% CIs are entirely within the prespecified range of equivalence, in our case defined as between –10 and +10.
In the left trial 95% CIs are between –9.5 and + 19.5, and thus the null hypothesis of no equivalence cannot be rejected. In the middle trial 95% CI are between -1.3 and 11.3, while in the right trial 95% CI are between -3.3 and 6.7 . This means that the last trial has a positive outcome: equivalence is demonstrated, the null hypothesis of no equivalence can be rejected. The negative correlation trial and the zero correlation trial despite a small mean difference between the two treatments, are not sensitive to reject the null-hypothesis, and this is obviously so because of the wide confidence intervals associated with negative and zero correlations.

10. RANK TESTING

Non-parametric tests are an alternative for ANOVA or t-tests when the data do not have a normal distribution. In that case the former tests are more sensitive than the latter. They are quick and easy, and are based on ranking of data in their order of magnitude. With heavily skewed data this means that we make the distribution of the ranks look a little bit like a normal distribution. We have paired and unpaired non-parametric tests and with the paired test the same problem of loss of sensitivity with negative correlations is encountered as the one we observed with the paired normality tests as discussed in the preceding paragraph. Non-parametric tests are also used to test normal distributions, and provide hardly different results from their parametric counterparts when distributions are approximately normal. Most frequently used tests:

For paired comparisons:

Wilcoxon signed rank test= **paired Wilcoxon** test

For unpaired comparisons:

Mann-Whitney test **= Wilcoxon rank sum** test

PAIRED TEST: WILCOXON SIGNED RANK TEST

Table 11. Paired comparison using Wilcoxon signed rank test: placebo-controlled clinical trial to test efficacy of sleeping drug

	Hours of sleep			rank
Patient	drug	placebo	difference	(ignoring sign)
1	6.1	5.2	0.9	3.5[x]
2	7.0	7.9	-0.9	3.5
3.	8.2	3.9	4.3	10
4.	7.6	4.7	2.9	7
5.	6.5	5.3	1.2	5
6.	8.4	5.4	3.0	8
7.	6.9	4.2	2.7	6
8.	6.7	6.1	0.6	2
9.	7.4	3.8	3.6	9
10.	5.8	6.3	-0.5	1

[x]number 3 and 4 in the rank are tight, so we use 3.5 for both of them.

The Wilcoxon signed rank test uses the signs and the relative magnitudes of the data instead of the actual data (Table 11). E.g., the above table shows the number of hours sleep in 10 patients tested twice: with sleeping pill and with placebo. We have 3 steps:

1. exclude the differences that are zero, put the remaining differences in ascending order of magnitude and ignore their sign and give them a rank number 1, 2, 3 etc (if differences are equal, average their rank numbers: 3 and 4 become 3.5 and 3.5);
2. add up the positive differences as well as the negative differences;
   + ranknumbers= 3.5+10+7+5+8+6+2+9=50.5
   - ranknumbers= 3.5+1= 4.5
3. The null hypothesis is that there is no difference between += and- ranknumbers. We assess the smaller of the two ranknumbers. The test is significant if the value is smaller than could be expected by chance. We consult the Wilcoxon signed rank table showing us the upper values for 5%, and 1% significance, for the number of differences constituting our rank. In this example we have 10 ranks: 5% and 1% points are respectively 8 and 3. The result is significant at $P<0.05$, indicating that the sleeping drug is more effective than the placebo.

UNPAIRED TEST: MANN-WHITNEY TEST

Table 12 shows two-samples of patients are treated with 2 different NSAID agents. Outcome variable is plasma globulin concentration (g/l). Sample one is printed in standard and sample 2 is printed in fat print.

Table 12. Two-samples of patients are treated with 2 different NSAIDs. Outcome variable is plasma globulin concentration (g/l). Sample one is printed in standard and sample 2 is printed in fat print

Globulin concentration(g/l)	ranknumber
26	1
27	**2**
28	**3**
29	4
30	5
31	6
32	7
33	8
34	**9**
35	10
36	11
38	12.5
38	**12.5**
39	**14.5**
39	**14.5**
40	**16**
41	17
42	**18**
45	**19.5**
45	**19.5**

We have 2 steps (Table 12):

1. The data from both samples are ranked together in ascending order of magnitude. Equal values are averaged.
2. Add up the rank numbers of each of the two samples. In sample-one we have 81.5, in sample-two we have 128.5. We now can consult the Table for Mann-Whitney tests and find with n=10 and n=10 (differences in sample sizes are no problem) that the smaller of the two sums of ranks should be smaller than 71 in order to conclude $P<0.05$. We can therefore not reject the null hypothesis of no difference, and have to conclude that the two samples are not significantly different from each other.

11. RANK TESTING FOR 3 OR MORE SAMPLES

THE FRIEDMAN TEST FOR PAIRED OBSERVATIONS

Table 13. Paired comparison to test efficacy of 2 dosages of a sleeping drug versus placebo on hours of sleep

	Hours of sleep					
Patient	dose 1 (hours)	dose 2	placebo	dose 1 (ranks)	dose 2	placebo
1	6.1	6.8	5.2	2	3	1
2	7.0	7.0	7.9	1.5	1.5	3
3.	8.2	9.0	3.9	2	3	1
4.	7.6	7.8	4.7	2	3	1
5.	6.5	6.6	5.3	2	3	1
6.	8.4	8.0	5.4	3	2	1
7.	6.9	7.3	4.2	2	3	1
8.	6.7	7.0	6.1	2	3	1
9.	7.4	7.5	3.8	2	3	1
10.	5.8	5.8	6.3	1.5	1.5	3

The Friedman test is used for comparing three or more repeated measures that are not normally distributed, and is an extension of the Wilcoxon signed rank test. An example is given in Table 13. The data are ranked for each patient in ascending order of hours of sleep. If the hours are equal, then an average ranknumber is given. Then, for each treatment the squared ranksum is calculated: for dose 1 it equals $(2+1.5+2+2+2+3+2+2+2+1.5)^2 = 400$, for dose 2 it is 676, for placebo it is 196. The following equation is used:

$$\text{chi-square} = \frac{12}{nk\,(k+1)} \left(\text{ranksum}_{\text{dose1}}{}^2 + \text{ranksum}_{\text{dose2}}{}^2 + \text{ranksum}_{\text{placebo}}{}^2\right) - 3n(k+1)$$

where n = the number of patients and k = the number of treatments.

The chi-square value is calculated to be 7.2. The chi-square statistic will be addressed in Chapter 3. briefly,it works very similar to the t-statistics. Chi-square values larger than the ones given in the chi-square table in the Appendix indicate that the null-hypothesis of no difference in the data can be rejected. In this example the calculated chi-square value is larger than the rejection chi-square for (3-1) degrees of freedom at $p = 0.05$, and, therefore, we conclude that there is a significant difference between the three treatments at $p < 0.05$. Post-hoc subgroups analyses (using Wilcoxon's tests) are required to find out exactly where the difference is situated, between group 1 and 2, between group 1 and 3, or between

group 2 and 3 or between two or more groups. The subject of post-hoc testing will be further discussed in Chapter 8.

THE KRUSKALL-WALLIS TEST FOR UNPAIRED OBSERVATIONS

Table 14. Three-samples of patients are treated with placebo or 2 different NSAIDs. The outcome variable is the fall in plasma globulin concentration (g/l). Group 1 patients are printed in italics, group 2 in normal standard and group 3 in fat standard print

Globulin concentration(g/l)	ranknumber
-17	***1***
-16	***2***
-5	***3***
-3	***4***
-2	***5***
16	***6***
18	***7***
26	8
27	**9**
28	***10.5***
28	**10.5**
29	12
30	14
30	***14***
30	***14***
31	16
32	17
33	18
34	**19**
35	20
36	21
38	22.5
38	**22.5**
39	**24.5**
39	**24.5**
40	**26**
41	27
42	**28**
45	**29.5**
45	**29.5**

The Kruskall-Wallis test compares multiple groups that are unpaired and not normally distributed, and is an extension of the Mann-Whitney test. Three groups of patients with rheumatoid arthritis are treated with a placebo or one of two different NSAIDS (Table 14). The fall in plasma globulin (g/l) is used to estimate the effect of treatments. First, we give a ranknumber to every patient dependent on his/her magnitude of fall. If two or three patients have the dame fall, they are given an average ranknumber. Then, we calculate the sum of the ranks for the three groups. For group 1 this amounts to 1+2+3+4+5+6+7+10.5+14+14 = 66.5, for group 2 to 175.5, group 3 to 488.5. Then we use the equation:

$$\text{chi-square} = \frac{12}{30\,(30-1)}\left(\frac{\text{ranksum}_{\text{group1}}^{\,2}}{10} + \frac{\text{ranksum}_{\text{group2}}^{\,2}}{10} + \frac{\text{ranksum}_{\text{group3}}^{\,2}}{10}\right) - 3(30-1)$$

where the number 30 equals all values, 10 the patient number per group.

The chi-square equals 7744.3. The chi-square statistic will be further addressed in chapter 3. It works very similar to the t-statistics. Briefly, chi-square values larger than the ones given in the chi-square table in the Appendix indicate that the null-hypothesis of no difference in the data can be rejected. In this example the calculated chi-square value is much larger than the rejection chi-square for (3-1) degrees of freedom and, therefore, we conclude that there is a significant difference between the three treatments at $p < 0.001$. Post-hoc subgroups analyses (using Man-Whitney tests) are required to find out exactly where the difference is situated, between group 1 and 2, between group 1 and 3, or between group 2 and 3 or between two or more groups. The subject post-hoc testing will be further discussed in Chapter 8.

12. CONCLUSIONS

For the analysis of efficacy data we test null-hypotheses. The t-test is appropriate for two parallel-groups or two paired samples. Analysis of variance (ANOVA) is appropriate for analyzing more than two groups / treatments. For data that do not follow a normal frequency distribution non-parametric tests are available: for paired data the Wilcoxon signed rank or Friedman tests, for unpaired data the Mann-Whitney test or Kruskall-Wallis tests are adequate.
Note: in the references an overview of relevant textbooks on the above subjects is given.

13. REFERENCES

1. Field A. Discovering Statistics using SPSSD. 2nd Edition, Sage Publications, Thousand Oaks, CA, USA, 2005.

2. Matthews DE, Farewell VT. Using and understanding medical statistics. Karger, Melbourne, Australia, 1996.
3. Cohen A, Posner J. Clinical Drug Research. Kluwer Academic Publishers, Dordrecht, Netherlands, 1995.
4. Bailar JC, Mosteller F. Medical Uses of Statistics. N Engl J Med Books, Waltham, MA, 1986.
5. Swinscow TD. Statistics at square one. BMJ Publishing Group, London, UK, 1996.
6. Glantz SA. Primer of Biostatistics. McGraw-Hill, Singapore, 1992.
7. Motulsky H. Intuitive Statistics. Oxford University Press. New York, 1995.
8. Kuhlmann J, Mrozikiewicz A. What should a clinical pharmacologist know to start a clinical trial? Zuckschwerdt Verlag, Munich, 1998.
9. De Vocht A. SPSS basic guide book. Bijleveld Press, Amsterdam, 1998.
10. Hays WL. Statistics. Holt, Rine and Winston, Toronto, Ontario, 1988.
11. Kirkwood BR. Medical statistics. Blackwell Scientific Publications, Boston, MA, 1990.
12. Petrie A, Sabin C. Medical Statistics at a Glance. Blackwell Science, London, UK, 2000.
13. Riffenburgh RH. Statistics in Medicine. Academic Press. New York, USA, 1999.
14. Utts JM. Seeing through Statistics. Brooks Cole Company, Pacific Grove, CA, 1999.
15. Glaser AN. High Yield Statistics. Lippincott, Williams & Wilkins Baltimore, Maryland, 2001.
16. Petrie A, Sabin C. Medical Statistics at a Glance. Blackwell Science. Malden, MA, 2000.
17. Lu Y, Fang JQ. Advanced Medical Statistics. World Scientific, River Edge, NJ, 2003.
18. Riegelman RK. Studying a study and testing a test. Lippincott Williams & Wilkins, 5th Edition, Philadelphia, 2005.
19. Peat J, Barton B. Medical statistics. Fifth Edition. BMJ Books, Blackwell Publishing, New Delhi,2005.
20. Campbell MJ. Statistics at square two. BMJ Books, Blackwell Publishing, New Delhi,2006.

CHAPTER 3

THE ANALYSIS OF SAFETY DATA

1. INTRODUCTION, SUMMARY DISPLAY

As discussed in chapter 1 the primary object of clinical trials of new drugs is generally to demonstrate efficacy rather than safety. However, a trial in human beings not at the same time adequately addressing safety is unethical, and the assessment of safety variables is an important element of the trial.
An effective approach to the analysis of adverse effects is to present summaries of prevalences. We give an example (table 1). Calculations of the 95% confidence intervals (CIs) of a proportion are demonstrated in Chapter 1. If 0.1<proportion (p) <0.9, then the binomial distribution is very close to the normal distribution, but if p < 0.1, the data follow a skewed, otherwise called Poisson distribution. 95 % CIs are, then, more adequately calculated according to $\pm 1.96\sqrt{p/n}$ rather than $\pm 1.96\sqrt{p(1-p)/n}$ (confer page 10). Alternatively, tables (e.g., Wissenschaftliche Tabelle, Documenta Geigy, Basel, 1995) and numerous statistical software packages can readily provide you with the CIs.

Table 1. The prevalence of side-effects after 8 week treatment

	Alpha blocker n=16			Beta blocker n=15		
side effect	yes	no	95% CIs(%)	yes	no	95% CIs (%)
nasal congestion	10	6	35-85	10	5	38-88
alcohol intolerance	2	12	2-43	2	13	4-71
urine incontinence	5	11	11-59	5	10	12-62
disturbed ejaculation	4	2	22-96	2	2	7-93
disturbed potence	4	2	22-96	2	2	7-93
dry mouth	8	8	25-75	11	4	45-92
tiredness	9	7	30-80	11	4	45-92
palpitations	5	11	11-59	2	13	2-40
dizziness at rest	4	12	7-52	5	10	12-62
dizziness with exercise	8	8	25-75	12	3	52-96
orthostatic dizziness	8	8	25-75	10	5	38-88
sleepiness	5	10	12-62	9	6	32-84

Table 1 gives an example. The numbers in the table relate to the numbers of patients showing a particular side effect. Some questions were not answered by all patients. Particularly, sleepiness occurred differently in the two groups: 33% in the

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 45–62.

left, 60% in the right group. This difference may be true or due to chance. In order to estimate the size of probability that this difference occurred merely by chance we can perform a statistical test which in case of proportions such as here has to be a chi-square or given the small data a Fisher exact test. We should add at this point that although mortality/morbidity may be an adverse event in many trials, there are also trials that use them as primary variables. This is particularly so with mortality trials in oncology and cardiology research. For the analysis of these kinds of trials the underneath methods of assessments are also adequate.

2. FOUR METHODS TO ANALYZE TWO UNPAIRED PROPORTIONS

Many methods exists to analyze two unpaired proportions, like odds ratios analysis (this chapter) and logistic regression (chapter 14), but here we will start by presenting the four most common methods for that purpose. Using the sleepiness data from above we construct a 2x2 contingency table:

	Sleepiness	no sleepiness
Left treatment (left group)	5 (a)	10 (b)
Right treatment (right group)	9 (c)	6 (d)

Method 1

We can test significance of difference similarly to the method used for testing continuous data (chapter 2). In order to do so we first have to find the standard deviation (SD) of a proportion. The SD of a proportion is given by the formula $\sqrt{p(1-p)}$. Unlike the SD for continuous data (see formula chapter 1), it is strictly independent of the sample size. It is not easy to prove why this formula is correct. However, it may be close to the truth considering an example (figure 1). Many samples of 15 patients are assessed for sleepiness. The proportion of sleepy people in the population is 10 out of every 15. Thus, in a representative sample from this population 10 sleepy patients will be the number most frequently encountered. It also is the mean proportion, and left and right from this mean proportion proportions grow gradually smaller, according to a binomial distribution (which becomes normal distribution with large samples). Figure 1 shows that the chance of 8 or fewer sleepy patients is 15% (area under the curve, AUC, left from 8.3 = 15%). The chance of 6 or less sleepy patients is 2.5 % (AUC left from 6.6 = 2.5%). The chance of 5 or less sleepy patients = 1%. This is a so-called binomial frequency distribution with mean 10 and a standard deviation of p (1-p)= 10/15 (1-5/15)= 1.7. -1SD means AUC of approximately 15%, -2SDs means AUC of approximately 2.5%. And, so, according to the curve below SD= p(1-p) is close to the truth.

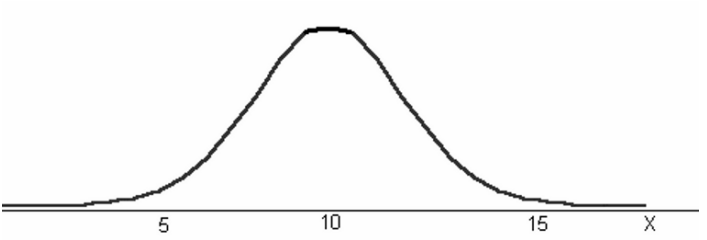


Figure 1. Frequency distribution of numbers of sleepy people observed in multiple samples of 15 patients from the same population.

Note: for null-hypothesis-testing standard error (SE) rather than SD is required, and SE = SD/ √n.
For testing we use the normal test (= z-test for binomial or binary data) which looks very much like the T-test for continuous data. T= d/SE , z= d/SE, where d= mean difference between two groups or difference of proportions and SE is the pooled SE of this difference. What we test is, whether this ratio is larger than approximately 2 (1.96 for proportions, a little bit more, e.g., 2.1 or so, for continuous data).

Example of *continuous* data (testing two means).

	Mean ± SD			$SEM^2 = SD^2 /n$
group 1 (n=10)	5.9	± 2.4	liter/min	5.76/10
group 2 (n=10)	4.5	± 1.7	liter/min	2.89/10

Calculate: $mean_1$ - $mean_2$ = 1.4.
Then calculate pooled $SEM = \sqrt{(SEM_1^2 + SEM_2^2)} = 0.930$.
Note: for SEM of difference: take square root of sums of squares of separate SEMs and, so, reduce the analysis of two means to one of a single mean.

$T = \frac{mean_1 - mean_2}{Pooled SEM} = 1.4/0.930 = 1.505$, with degrees of freedom (dfs) 18* p>0.05.
*We have 2 groups of n=10 which means 2 x 10 -2 = 18 dfs.

Example of *proportional* data (testing two proportions).

2x2 table	Sleepiness	No sleepiness
Left treatment (left group)	5	10
Right treatment (right group)	9	6

$$z = \frac{\text{difference between proportions of sleepers per group (d)}}{\text{pooled standard error difference}}$$

$$z = \frac{d}{\text{pooled SE}} = \frac{(9/15 - 5/15)}{\sqrt{(SE_1^2 + SE_2^2)}}$$

$$SE_1 \text{ (or } SEM_1 \text{)} = \sqrt{\frac{p_1 (1 - p_1)}{n_1}} \text{ where } p_1 = 5/15 \text{ etc.....,}$$

z = 1.45, not statistically significant from zero, because for a $p<0.05$ a z-value of at least 1.96 is required.

Note: the z-test uses the bottom row of the t-table (see APPENDIX), because, unlike continuous data that follow a t-distribution, proportional data follow a normal distribution. The z-test is improved by inclusion of a continuity correction. For that purpose the term - $(1/2n_1 + 1/2n_2)$ is added to the denominator where n_1 and n_2 are the sample sizes. The reason is that a continuous distribution is used to approximate a proportional distribution which is discrete, in this case binomial.

Methods 2

According to some a more easy way to analyze proportional data is the chi-square test. The chi-square test assumes that the data follow a chi-square frequency distribution which can be considered the square of a normal distribution (see also chapter 22). First some philosophical considerations.
Repeated observations have both (1) a central tendency, and (2) a tendency to depart from an expected overall value, often the mean. In order to make predictions an index is needed to estimate the departures from the mean. Why not simply add up departures? However, this doesn't work, because, with normal frequency distributions, the add-up sum is equal to 0. A pragmatic solution chosen is taking the add-up sum of $(\text{departures})^2$ = the variance of a data sample. Means / proportions follow normal frequency distributions, variances follow **$(\text{normal-distribution})^2$**. The normal distribution is a biological rule used for making predictions from random samples.
With a normal frequency distribution in your data (Figure 2 upper graph) you can test whether the mean of your study is significantly different from 0.
If the mean result of your study > approximately 2 SEMs distant from 0, then we have <5% chance of no difference from 0, and we are entitled to reject the 0-hypothesis of no difference.
With $(\text{normal frequency distributions})^2$ (Figure 2 lower graph) we can test whether the variance of our study is significantly different from 0. If the variance of our study is $> 1.96^2$ distant from 0, then we have <5% chance of no difference from 0, and we are entitled to reject the 0-hypothesis of no difference.

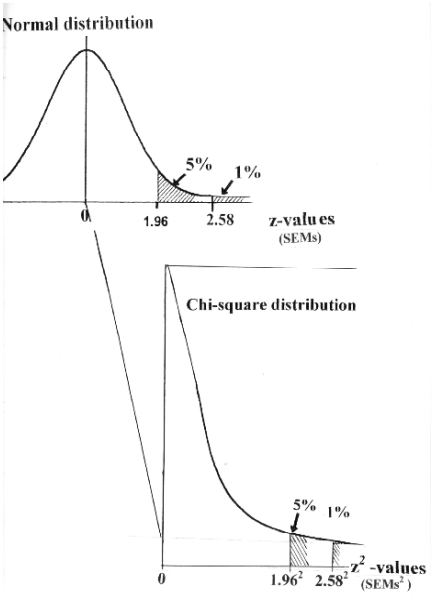


Figure 2. Normal and chi-square frequency distributions.

The chi-square test, otherwise called χ^2 test can be used for the analysis of two unpaired proportions (2x2 table), but first we give a simpler example, a 1x2 table

Sleepy	Not-sleepy	Sleepy	Not-sleepy
observed (O)		expected from population (E)	
a (n=5)	b (n=10)	α (n=10)	β (n=5)

We wish to assess whether the observed proportion is significantly different from the established population data from this population, called the expected proportion?

O –E =

a- α = 5-10 = -5

b- β = 10-5 = 5 +

0 doesn't work

The above method to assess a possible difference between the observed and expected data does not work. Instead, we take square values.

$(a- \alpha)^2 = 25$ divide by α to standardize = 2.5

$(b- \beta)^2 = 25$ " " β " " = 5 +

7.5

χ^2 Value = the add-up variance in data = 7.5

α is the standard error (SE) of $(a- \alpha)^2$ and is used to standardize the data, similarly to the standardization of mean results using the t-statistic (replacing the mean results with t-values, see Chapter 2).

This 1x2 table has 1 degree of freedom. The chi-square table (see APPENDIX) shows four columns of chi-square values (standardized variances of various studies), an upper row of areas under the curve (AUCs), and a left end column with the degrees of freedom. For finding the appropriate area under the curve (= p-value) of a 1x2 table we need the second row, because it has 1 degree of freedom. A chi-square value of 7.5 means an AUC = p-value of <0.01. The O-hypothesis can be rejected. Our observed proportion is significantly different from the expected proportion.
Slightly more complex is the chi-square test for the underneath table of observed numbers of patients in a random sample:

	Sleepiness(n)	no sleepiness(n)
Left treatment (left group)	5 (a)	10 (b)
Right treatment (right group)	9 (c)	6 (d)

n = numbers of patients in each cell.

Commonly, no information is given about the numbers of patients to be expected, and, so, we have to use the best estimate based of the data given. The following procedure is applied:

cell a: $(O-E)^2 / E = (5 - 14/30 \times 15)^2 / 14/30 \times 15 = ..$
" b: $(O-E)^2 / E$
" c: $(O-E)^2 / E$
" d: $(O-E)^2 / E$ +
chi-square = 2.106

(O= observed number; E= expected number=(proportion sleepers /total number) x $number_{group}$).

We can reject the 0-hypothesis if the squared distances from expectation > $(1.96)^2$ = 3.841 distant from 0, which is our critical chi-square value required to reject the 0-hypothesis. A chi-square value of only 2.106 means that the 0-hypothesis can not be rejected.
Note: a chi-square distribution = a squared normal distribution. When using the chi-square table, both the 1x2 and the 2x2 contingency tables have only 1 degree of freedom.

Method 3

Instead of the above calculations to find the chi-square value for a 2x2 contingency table, a simpler pocket calculator method producing exactly the same results is described underneath

	Sleepiness	no sleepiness	total
Left treatment (left group)	5 (a)	10 (b)	a+b
Right treatment (right group)	9 (c)	6 (d)	c+d
	a+c	b+d	

Calculating the chi-square (χ^2)- value is calculated according to:

$$\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

In our case the size of the chi-square is again 2.106 at 1 degree of freedom which means that the 0-hypothesis of no difference not be rejected. There is no significant difference between the two groups.

Method 4

Fisher-exact test is used as contrast test for the chi-square or normal test, and also for small samples, e.g., samples of n < 100. It, essentially, makes use of faculties expressed as the sign **"!"**: e.g., 5 **!** indicates 5x4x3x2x1.

	Sleepiness	no sleepiness
Left treatment (left group)	5 (a)	10 (b)
Right treatment (right group)	9 (c)	6 (d)

$$P= \frac{(a+b)!\,((c+d)!\,(a+c)!\,(b+d)!}{(a+b+c+d)!\,a!b!c!d!} = 0.2 \quad \text{(much larger than 0.05)}$$

Again, we can not reject the null-hypothesis of no difference between the two groups. This test is laborious but a computer can calculate wide faculties in seconds.

3. CHI-SQUARE TO ANALYZE MORE THAN TWO UNPAIRED PROPORTIONS

As will be explained in chapter 23, with chi-square statistics we enter the real world of statistics, because it is used for multiple tables, and it is also the basis of analysis of variance. Large tables of proportional data are more frequently used in business statistics than they are in biomedical research. After all, clinical investigators are, generally, more interested in the comparison between two treatment modalities than they are in multiple comparisons. Yet, e.g., in phase 1 trials multiple compounds are often tested simultaneously. The analysis of large tables is similar to that of the above method-2. For example:

	Sleepiness	no sleepiness
Group I	5 (a)	10 (b)
Group II	9 (c)	6 (d)
Group III	.. (e)	... (f)
Group IV	..	
Group V		

cell a: $(O-E)^2 / E =$
b: $(O-E)^2 / E$
c: $(O-E)^2 / E$
d: $(O-E)^2 / E$
e: ..
f : ..

chi-square value = .. +

For cell a O = 5

$$E = \frac{(5+9+...)}{(5+10+9+6+...)} x(5+10) \quad \text{etc}$$

Large tables have many degrees of freedom (dfs). For 2x2 cells, we have (2-1) x (2-1) = 1df, 5% p-value at chi-square = 3.841. For 3x2 cells, we have (3-1) x (2-1) = 2dfs, 5% p-value at chi-square = 5.991. For 5x2 cells, we have (5-1)(2-1) = 4 dfs, 5% p-value at chi-square = 9.488. Each degree of freedom has its own frequency distribution curve (Figure 3):

dfs 2=>p=0.05 at χ^2 5.99
dfs 4 p=0.05 at χ^2 9.49
dfs 6 p=0.05 at χ^2 12.59
dfs 8 p=0.05 at χ^2 15.51
dfs 10 p=0.05 at χ^2 18.31.

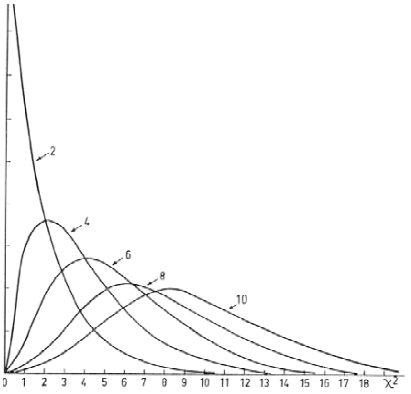


Figure 3. Each degree of freedom has its own frequency distribution curve.

As an example we give a χ^2 test for 3x2 table

Hypertension	yes	no
Group 1	a n = 60	d n = 40
Group 2	b n = 100	e n = 120
Group 3	c n = 80	f n = 60

Give the best estimate of the expected numbers in the cell according to the method described for the 2x2 contingency table above. Per cell: divide hypertensives in study by observations in study, multiply by observations in group. It gives you the best estimate. For cell a this is α =[(a+b+c) / (a+b+c+d+e+f)] x (a+d). Do the same for each cell and add-up:

α = [(a+b+c) / (a+b+c+d+e+f)] x (a+d) = 52.17
β ..... =114.78
γ = 73.04
δ = [(d+e+f))/(a+b+c+d+e+f)] x (a+d) = 47.83
ε .... = 57.39
ξ .... = 66.96

$(a-\alpha)^2/\alpha$ = 1.175
(b- ... = 1.903
(c- ... = 0.663
(d- ... = 1.282
(e- ... = 68.305
(f- ... = 0.723 +
χ^2 value = 72.769

The p-value for (3-1)x(2-1) = 3 degrees of freedom is <0.001 according to the chi-square table (see APPENDIX).

Another example is given, a 2x3 table:

Hypertension	hypertens-yes /	hypertens-no /	don't know
Group 1	(a) n=60	(c) n= 40	(e) n=60
Group 2	(b) n=50	(d) n= 60	(f) n=50

Give best estimate population. Per cell: divide hypertensives in population by all patients, multiply by hypertensives in group. For cell a this is:

$\alpha = [(a+b) / (a+b+c+d+e+f)] \times (a+c+e)$

Calculate every cell, add-up results.

$\alpha = [(a+b) / (a+b+c+d+e+f)] \times (a+c+e) = 55.000$
β .... $= 55.000$
$\gamma = [(c+d) / (a+b+c+d+e+f)] \times (a+c+e) = 51.613$
δ = ... $= 51.613$
ε ... $= 55$
ξ ... $= 55$

$(O-E)^2 / E =$
$(a-\alpha)^2 / \alpha = 0.45$
(b ... $= 0.45$
(c .. $= 0.847$
(d .. $= 1.363$
(e .. $= 0.45$
(f .. $= 0.45$ +

$\chi^2 = 4.01$

For (2-1) x (3-1) = 2 degrees of freedom our p-value is <0.001 according to the chi-square table (see APPENDIX).

4. MCNEMAR'S TEST FOR PAIRED PROPORTIONS

Paired proportions have to be assessed when e.g. different diagnostic tests are performed in one subject. E.g., 315 subjects are tested for hypertension using both an automated device (test-1) and a sphygmomanometer (test-2), (Table 2).

Table 2. Finding discordant pairs

		Test 1		
		+	-	total
Test 2	+	184	54	238
	-	14	63	77
total		198	117	315

$$\text{Chi - square McNemar} = \frac{(54-14)^2}{54+14} = 23.5$$

184 subjects scored positive with both tests and 63 scored negative with both tests. These 247 subjects therefore give us no information about which of the tests is more likely to score positive. The information we require is entirely contained in the 68 subjects for whom the tests did not agree (the discordant pairs). Table 2 shows how the chi-square value is calculated. Here we have again 1 degree of freedom, and so, a chi-square value of 23.5 indicates that the two devised produce significantly different results at $p<0.001$.

To analyze samples of more than 2 pairs of data, e.g., 3, 4 pairs, etc., McNemar's test can not be applied. For that purpose Cochran's test or logistic regression analysis is adequate (chapter 14).

5. SURVIVAL ANALYSIS

Survival analysis

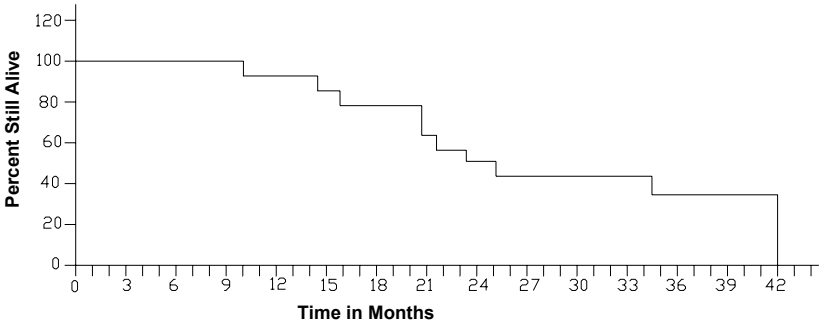


Figure 4. Example of a survival curve plotting survival as a function of time.

A survival curve plots percentage survival as a function of time. Figure 4 is an example. Fifteen patients are followed for 36 months. At time zero everybody is alive. At the end 40% (6/15) patients are still alive. Percentage decreased whenever a patient died. A problem with survival analysis generally is that of lost data: some patients may be still alive at the end of the study but were lost for follow-up for several reasons. We at least know that they lived at the time they were lost, and so they contribute useful information. The data from subjects leaving the study are called **censored** data and should be included in the analysis.

With the **Kaplan-Meier** method, survival is recalculated every time a patient dies (approaches to survival different from the Kaplan-Meier approach are (1) the actuarial method, where the x-axis is divided into regular intervals and (2) life-table analysis using tables instead of graphs). To calculate the fraction of patients who survive a particular day, simply divide the numbers still alive after the day by the number alive before the day. Also exclude those who are lost=censored on the very day and remove from both the numerator and denominator. To calculate the fraction of patients who survive from day 0 until a particular day, multiply the fraction who survive day-1, times the fraction of those who survive day-2, etc. This product of many survival fractions is called the **product-limit**. In order to calculate the 95% CIs, we can use the formula:

95% CI of the product of survival fractions (p) at time $k = p \pm 2 \cdot p\sqrt{\frac{(1-p)}{k}}$

The interpretation: we have measured survival in one sample, and the 95%CI shows we can be 95% sure that the true population survival is within the boundaries (see figure upper and lower boundaries). Instead of days, as time variable, weeks, months etc may be used.

Testing significance of difference between two Kaplan-Meier curves

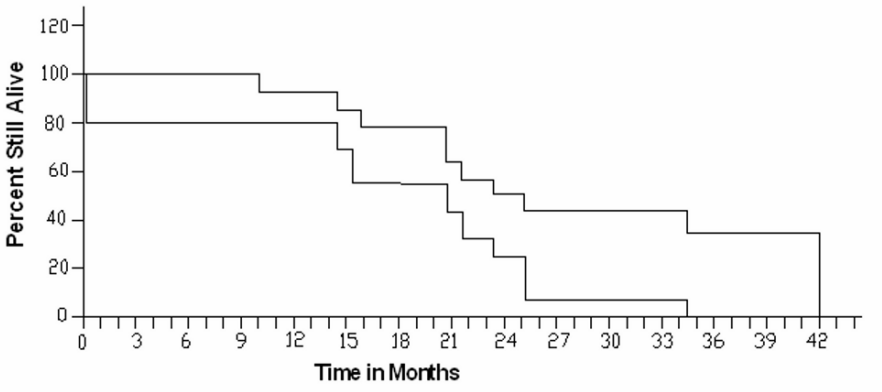


Figure 5. Two Kaplan-Meier survival curves.

Survival is essentially expressed in the form of either proportions or odds, and statistical testing whether one treatment modality scores better than the other in terms of providing better survival can be effectively done by using tests similar to the above **chi-square tests** or chi-square-like tests in order to test whether any proportion of responders is different from another proportion, e.g., the proportion of responders in a control group. RRs or ORs are calculated for that purpose (review chapter 1). For example, in the example in the i-th 2-month period we have left the following numbers: a_i and b_i in curve 1, c_i and d_i in curve 2,

Contingency table	Numbers of deaths	numbers alive
Curve 1	a_i	b_i
curve 2	c_i	d_i

i = 1,2,3,...

$$\text{Odds ratio} = \frac{a_i/b_i}{c_i/d_i} = \frac{a_i d_i}{b_i c_i}$$

Significance of difference between the curves (Figure 5) is calculated according to the added products "ad" divided by "bc". This can be readily carried out by the

Mantel-Haenszl summary chi-square test:

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

where we thus have multiple 2x2 contingency tables e.g. one for every last day of a subsequent month of the study. With 18 months follow-up the procedure would yield 18 2x2-contingency-tables. This Mantel Haenszl summary chi square test is, when used for comparing survival curves, more routinely called **log rank test** (this name is rather confusing because there is no logarithm involved)

Note: An alternative more sophisticated approach to compare survival curves is the **Cox's proportional hazards model**, a method analogous to **multiple regression analysis** for multiple means of continuous data and to **logistic regression** for proportions in a multivariate model (chapter 15).

6. ODDS RATIO METHOD FOR ANALYZING TWO UNPAIRED PROPORTIONS

Odds ratios increasingly replace chi/square tests for analyzing 2x2 contingency tables.

	illness	no illness
group 1	a	b
group 2	c	d

The odds ratio (OR)= a/b / c/d
= odds of illness group1 / odds illness group 2
= chance illness...../.....

We want to test whether the OR is significantly different from an OR of 1.0.
For that purpose we have to use the logarithmic transformation, and so we will start by recapitulating the principles of logarithmetic calculations.
Log = log to the base 10; Ln = natural log = log to the base e (e=2.71…)

log 10= ^{10}log 10= 1
log 100= ^{10}log 100= 2
log 1= ^{10}log 1= ^{10}log 10^0 = 0
antilog 1= 10
antilog 2= 100
antilog 0 = 1

ln e= ^{e}log e = 1
ln e^2 = ^{e}log e^2 = 2
ln 1 = ^{e}log 1 = ^{e}log e^0 = 0
antiln 1= e
antiln 2= e^2
antiln 0= 1

The frequency distributions of samples of continuous numbers or proportions are normal. Those of many odds ratios are not. The underneath example is an argument that odds ratios may follow an exponential pattern, while the normal distribution has been approximated by mathematicians by means of the underneath exponential formula

$$\frac{a/b}{c/d}=\frac{1/10}{1/100}=10 \qquad \frac{a/b}{c/d}=\frac{1/10}{1/10}=1 \qquad \frac{a/b}{c/d}=\frac{1/100}{1/10}=\frac{1}{10}$$

$$y=\left(\frac{1}{\sqrt{2\pi}}\right)e^{\frac{-1}{2}x^2}$$

x individual data, y how often, e= 2.718.

It was astonishing but not unexpected that mathematicians discovered that frequency distributions of log OR followed a normal distribution, and that results were even better if ln instead of log was used.

	event	*no event*
group 1	*a*	*b*
group 2	*c*	*d*

If OR $=\frac{a/b}{c/d}=1$, this means that no difference exists between group 1 and 2.
If OR =1, then lnOR=0. With a *normal distribution* if the result > 2 standard errors (SEs) distant from 0, then the result is significantly different from 0 at p<0.05. This would also mean that, if ln OR > 2 SEs distant from 0, then this result would be significantly different from 0 at p<0.05. There are three possible situations:

```
study 1                 |  < --.-- >  lnOR > 2 SEs dist 0 =>  p<0.05
study 2           < --.-|-.-- >       lnOR < 2 SEs dist 0 =>  ns
study 3      < --.-- >  |             lnOR > 2 SEs dist 0 =>  p<0.05
             ...........|...........
                  lnOR=0 (OR=1.0)
```

Using this method we can test the OR. However, we need to know how to find the SE of our OR. SE of OR is given by the formula $\sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})}$.
This relatively simple formula is not a big surprise, considering that the SE of a number g = √g, and the SE of 1/g = $\sqrt{\frac{1}{g}}$. We can now assess our data by the OR method as follows:

	Hypertension yes	hypertension no
Group 1	a n = 5	b n=10
Group 2	c n=10	d n= 5

$$OR = \frac{a/b}{c/d} = 0.25$$

lnOR= -1.3863

$$\text{SEM lnOR} = \sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})} = 0.7746$$

ln OR ± 2 SEMs = -1.3863 ± 1.5182

= between– 2.905 and 0.132,

Now turn the ln numbers into real numbers by the antiln button of your pocket calculator.

= between 0.055 en 1.14.

The result "crosses" 1.0, and, so, it is not significantly different from 1.0.

A second example answers the question: is the difference between the underneath group 1 and 2 significant?

	orthostatic hypotension yes	no
Group 1	77	62
Group 2	103	46

$$OR = \frac{103/46}{77/62} = \frac{2.239}{1.242} = 1.803$$

lnOR = 0.589

$$\text{SEM lnOR} = \sqrt{(\frac{1}{103}+\frac{1}{46}+\frac{1}{77}+\frac{1}{62})} = 0.245$$

lnOR ± 2 SEMs = 0.589± 2(0.245)

= 0.589 ± 0.482

= between 0.107 and 1.071.

Turn the ln numbers into real numbers by use of antiln button of your pocket calculator.

= between 1.11 and 2.92 , and, so, significantly different from 1.0.

What p-value do we have: t=lnOR / SEM =0.589/0.245=2.4082. The bottom row of the t-table is used for proportional data (z-test), and give us a p-value <0.02.

Note: a major problem with odds ratios is the ceiling problem. If the control group n=0, then it is convenient to replace 0 with 0.5 in order to prevent this problem.

7. ODDS RATIOS FOR 1 GROUP, TWO TREATMENTS

So far we assessed 2 groups, 1 treatment. Now we will assess 1 group, 2 treatments and use for that purpose the **McNemar's OR**.

		normotension with drug 1	
		yes	no
normotension	yes	(a) 65	(b) 28
with drug 2	no	(c) 12	(d) 34

Here the OR = b/c, and the SE is not $\sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})}$, but rather $\sqrt{(\frac{1}{b}+\frac{1}{c})}$.

OR = 28/12 = 2.33
lnOR=ln2.33=0.847
SE= $\sqrt{(\frac{1}{b}+\frac{1}{c})}$ =0.345
lnOR ± 2 SE = 0.847 ± 0.690
= between 0.157 and 1.537,
Turn the ln numbers into real numbers by the anti-ln button of your pocket calculator.
= between 1.16 and 4.65
= sig diff from 1.0.

Calculation p-value: t = lnOR / SEM = 0.847 : 0.345 = 2.455. The bottom row of the t-table produces a p-value of <0.02, and the two drugs produce, thus, significantly different results at p<0.02.

8. CONCLUSIONS

1. For the analysis of efficacy data we test null-hypotheses, safety data consist of proportions, and require for statistical assessment different methods.
2. 2x2 tables are convenient to test differences between 2 proportions.
3. Use chi-square or t-test for normal distributions (z-test) for that purpose.
4. For paired proportions the McNemar's test is appropriate.
5. Kaplan Meier survival curves are also proportional data: include lost patients.
6. Two Kaplan-Meier Curves can be compared using the Mantel-Haenszl = Log rank test
7. Odds ratios with logarithmic transformation provide an alternative method for analyzing 2x2 tables.

In the past two chapters we discussed different statistical methods to test statistically experimental data from clinical trials. We did not emphasize

correlation and regression analysis. The point is that correlation and regression analysis test correlations, rather than causal relationships. Two samples may be strongly correlated e.g., two different diagnostic tests for assessment of the same phenomenon. This does, however, not mean that one diagnostic test causes the other. In testing the data from clinical trials we are mainly interested in causal relationships. When such assessments were statistically analyzed through correlation analyses mainly, we would probably be less convinced of a causal relationship than we are while using prospective hypothesis testing. So, this is the main reason we so far did not address correlation testing extensively. With epidemiological observational research things are essentially different: data are obtained from the observation of populations or the retrospective observation of patients selected because of a particular condition or illness. Conclusions are limited to the establishment of relationships, causal or not. We, currently, believe that relationships in medical research between a factor and an outcome can only be proven to be causal when between the factor is introduced and subsequently gives rise to the outcome. We are more convinced when such is tested in the form of a controlled clinical trial. A problem with multiple regression and logistic regression analysis as method for analyzing of multiple samples in clinical trials is closely related to this point. There is always an air of uncertainty about such regression data. Many trials use null-hypothesis testing of two variables, and use multiple regression data only to support and enhance the impact of the report, and to make readership more willing to read the report, rather than to prove the endpoints. It is very unsettling to realize that clinicians and clinical investigators often make bold statements about causalities from multivariate analyses. We believe that this point deserves full emphasis, and will, therefore, address it again in the Chapters 13 - 18.

CHAPTER 4

LOG LIKELIHOOD RATIO TESTS FOR SAFETY DATA ANALYSIS

1. INTRODUCTION

For Gandhi non-violence was a primary invariance principle, while for his political successor Nehru justice was so. Invariance principles signify that while everything changes in life, some laws of life do not. Consequently, these laws of life do not include a measure of error. For example, Einstein's invariance principle is expressed in the famous equation $E = mc^2$. Most statistical tests, including t - (and z-) tests, F-tests, chi-square tests, odds ratio tests, do not meet the invariance principle, because they apply *estimated* likelihoods like averages and proportions that have their standard errors as a measure of uncertainty. However, a few statistical tests use likelihoods without standard error. These tests, called exact tests, should, by their very nature, provide the best precision and sensitivity of testing. They include, among others, the Fisher exact test and the log likelihood ratio test. Particularly, the log likelihood ratio test, avoiding some of the numerical problems of the other exact likelihood tests, is straightforward, and is available through most major software programs[1-8], although infrequently used so far. This paper reviews the advantages and problems of the log likelihood ratio test, and gives real and hypothesized data examples supporting its better sensitivity. We do hope that the paper will stimulate researchers to more often apply this test.

2. NUMERICAL PROBLEMS WITH CALCULATING EXACT LIKELIHOODS

Proportions of patients with events are an important endpoint in cardiovascular research. They are traditionally analyzed in the form of a contingency table of four cells, otherwise called 2 x 2 contingency table, using chi-square tests or odds ratio tests.

	Number patients with events	number patients without
Group 1	a	b
Group 2	c	d

The problem with the traditional tests is that sensitivity is limited. As an alternative, the log likelihood ratio test, based on exact rather than estimated likelihoods, can be used. The general problem with exact likelihoods is, that they can be very complicated and may run into numerical problems that even modern computers can not handle. Let us assume that on average the proportion of patients with an event in a target population equals p. The likelihood of getting exactly y events in a sample of n individuals in this population can be calculated according to the underneath binomial equation:

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 63–71.

$$\text{Likelihood } p = \frac{n!}{y!(n-y)!} p^{y} (1-p)^{(n-y)}$$

n! = n faculty = n (n-1)(n-2)(n-3)…….

For example, a group of citizens was taking a pharmaceutical company to court for misrepresenting the danger of fatal rhabdomyolysis due to a statin treatment:

	Patients with rhabdomyolysis	patients without
company	1 (a)	309999 (b)
citizens	4 (c)	300289 (d)

p_{co} = proportion given by the pharmaceutical company = a / (a+b) = 1 / 310000
p_{ci} = proportion given by the citizens = c / (c+d) = 4 / 300293

$$\text{likelihood } p_{co} = \frac{310000!}{1!\,(310000-1)!} \cdot (1/310000)^{1} \cdot (1-1/310000)^{(310000-1)}$$

Likelihood p_{ci} can be calculated similarly.

The numerical problem of calculating likelihoods in the above way can be largely circumvented by taking the (log) ratios of two equations as will be demonstrated underneath.

3. THE NORMAL APPROXIMATION AND THE ANALYSIS OF CLINICAL EVENTS

If we take many samples from a target population, the mean results of those samples usually follow a normal frequency distribution, meaning that the value in the middle will be observed most frequently and the more distant from the middle the less frequently a value will be observed. E.g., we will have only 5% chance to find a result more than 2 standard errors (SEs) (or more precisely 1.96 SEs) distant from the middle. The same is true with proportional data like events. Many statistical tests make use of the normal distribution to make predictions. Figure 1 shows, e.g., how the normal distribution theorem is used to reject the null-hypothesis of no difference from zero.
Assume on average that 10 of 15 patients in a population will have some kind of cardiovascular event within a certain period of time. Then, 10/15 will be the proportion most frequently encountered when randomly sampling from this population. The chance of finding <10 or >10 gets gradually smaller. Figure 2 gives on the x-axis (often called z-axis in statistics) the results from many samples, the y-axis shows “how often”. The chance of 8 or less is only 15%, of 7 or less only 2.5 %, and of 5 or less only 1%. With many samples the graph follows a normal frequency distribution with 95% of the sample results between ± 2 SEs distant from the mean value, a proportion of 10/15. Most of the approaches to test

the significance of difference between the events in a treatment and control group make use of this normal approximation. This includes the z-test, the chi-square test, and the odds ratio test. Also, the log likelihood ratio test does so.

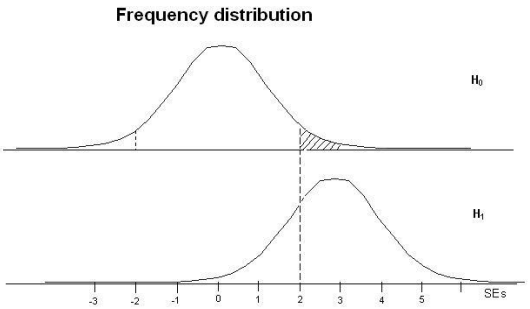


Fig. 1. H1 = graph based on the data of a sample with standard errors distant from zero (SEs) as unit on the x-axis, often called z-axis in statistics. H0 = same graph with a mean value of 0. We make a giant leap from the sample to the entire population, and we can do so because the sample is assumed to be representative for the entire population. H1 = also the summary of the means of many samples similar to our sample. H0 = also the summary of the means of many samples similar to our sample, but with an overall effect of 0. Our mean not 0 but 2.9. Still it could be an outlier of many samples with an overall effect of 0. If H0 is true, then our sample is an outlier. We can't prove, but calculate the chance/ probability of this possibility. A mean result of 2.9 SEs is far distant from 0: suppose it belongs to H0. Only 5% of H0 trials >2.0 SEs distant 0. The chance that it belongs to H0 is thus <5%. We conclude that we have <5% chance to find this result, and, therefore, reject this small chance.

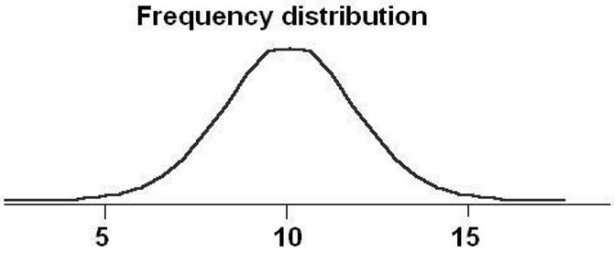


Fig. 2. Assume that, on average, 10 of 15 patients in a population will have some kind of cardiovascular event within a certain period of time. Then, 10/15 will be

the proportion most frequently encountered when taking many random samples of 15 patients from this population. The chance of finding <10 or >10 gets gradually smaller. On the x-axis the numbers of events from many samples is given, the y-axis shows "how often". The chance of 8 or less is only 15%, of 7 or less only 2.5 %, and of 5 or less only 1%. With many samples the graph follows a normal frequency distribution with 95% of the sample results between ± 2standard errors distant from the mean value.

4. LOG LIKELIHOOD RATIO TESTS AND THE QUADRATIC APPROXIMATION

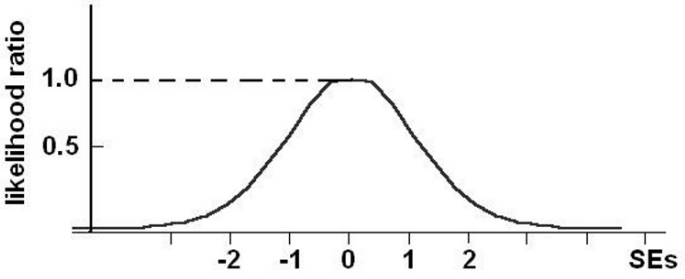


Fig. 3. Assume like in Figure 2 that 10 / 15 has the maximum likelihood, while all other proportions have less likelihood. The likelihood ratio is defined as the measured proportion / maximum likelihood. The likelihood ratio for 10 / 15 thus equals 1. If p = 10 / 15 is given place 0 on the z-axis with standard errors as unit, and the top of the curve = 1, then Figure 2 can also be interpreted as a likelihood ratio curve.

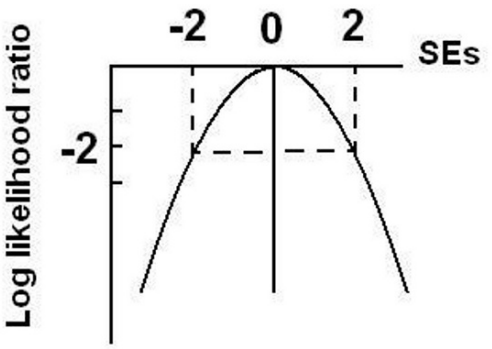


Fig. 4. If we transform the likelihood ratio values of the y-axis from Figure 3 to log likelihood ratio values, leaving the z-axis unchanged, then the above curve is observed.

Assume, like in the above example, that 10 / 15 has the maximum likelihood, while all other proportions have less than that. The likelihood ratio is defined as the measured proportion / maximum likelihood. The likelihood ratio for 10 / 15 thus equals 1. Instead of frequency distribution of many samples, Figure 2 can also be interpreted as a likelihood ratio curve of many samples. If p = 10 / 15 is given place 0 on the z-axis, with standard error-units on the z-axis and the top of the curve = 1, then the underneath normal distribution equation and the corresponding curve (Figure 3) are adequate.

$$\text{Likelihood ratio} = e^{-1/2\, z^2}$$

If we transform the likelihood ratio values of the y-axis from Figure 3 to log likelihood ratio values, leaving the z-axis unchanged, then the next equations and their corresponding curve (Figure 4) are adequate.

$$\text{log likelihood ratio} = -1/2\, z^2$$
$$-2 \text{ log likelihood ratio} = z^2$$

With normal distributions, if $z > 2$ or < -2, we conclude a significant difference from zero in the data at $p<0.05$. Here if -2 log likelihood ratio > 2 or < -2, then the difference between the proportions of events in a two-group comparison is significant at $p < 0.05$.
We now calculate the exact likelihoods for either of the two proportions using the above binomial equation.

$$\text{Likelihood } p = \frac{n!}{} \quad p^y \ (1-p)^{(n-y)}$$

$$\text{log likelihood } p = \log \frac{n!}{y!(n-y)!} + y \log p + (n-y) \log (1-p).$$

If the data produce two proportions, we can deduce from the above formula the exact (log) likelihood ratio of the two, where log is the natural logarithm. We take the previously used example.

	Patients with rhabdomyolysis	patients without
company	1 (a)	309999 (b)
citizens	4 (c)	300289 (d)

p_{co} = proportion given by the pharmaceutical company = a / (a+b) = 1 / 310000
p_{ci} = proportion given by the citizens = c / (c+d) = 4 / 300293

$$\text{log likelihood ratio} = \log \frac{\text{likelihood } p_{co}}{\text{likelihood } p_{ci}}$$
$$= \text{log likelihood } p_{co} - \text{log likelihood } p_{ci}$$
$$= y \log p_{co} / p_{ci} + (n-y) \log (1-p_{co})/ (1-p_{ci})$$

As -2 log likelihood ratio equals z^2 , we can now test the significance of difference between the two proportions.

$$\text{Log likelihood ratio} = 4 \log \frac{1/310000}{4/300293} + 300289 \log \frac{1- 1/310000}{1- 4/300293}$$
$$= -2.641199$$

-2 log likelihood ratio = 5.2824 ($p<0.05$, because $z > 2$).

We should note that both the odds ratio test and chi-square test produced a non-significant result here ($p>0.05$).

5. MORE EXAMPLES

Example 1
Two group of 15 patients at risk for arrhythmias were assessed for the development of torsade de points after calcium channel blockers treatment

	Patients with torsade de points	patients without
Calcium channel blocker 1	5	10
Calcium channel blocker 2	9	6

The proportion of patients with event from calcium channel blocker 1 is 5/15, from blocker 2 it is 9/15.

$$\text{Log likelihood ratio} = 9 \log \frac{5/15}{9/15} + 6 \log \frac{1-5/15}{1-9/15}$$
$$= -2.25$$

-2 log likelihood ratio = 4.50 ($p < 0.05$, because $z > 2$).

Both odds ratio test and chi-square test were again non-significant ($p>0.05$).

Example 2
Two groups of patients with stage IV New York Heart Association heart failure were assessed for hospitalizations after two beta-blockers.

	Patients with hospitalization	patients without
Beta blocker 1	77	62
Beta blocker 2	103	46

The proportion of patients with event from beta blocker 1 is 77 / 139, from beta blocker 2 it is 103 / 149.

$$\text{Log likelihood ratio} = 103 \log \frac{77/139}{103/149} + 46 \log \frac{1-77/139}{1-103/149}$$
$$= -5.882$$

-2 log likelihood ratio = 11.766 ($p < 0.002$, because $z > 3.090$).

Both the odds ratio test and chi-square test were also significant. However, at lower levels of significance, both p-values $0.01< p < 0.05$.

6. DISCUSSION

The chi-square test for events uses the observed cells in a contingency table to approximate the expected cells, a rather imprecise method. The odds ratio test uses the log transformation of a skewed frequency distribution as a rather imprecise approximation of the normal distribution. Sensitivity of these tests is, obviously, limited, and tests with potentially better sensitivity like exact tests are welcome.
At first sight, we might doubt about the precision of the log likelihood ratio test for events, because it is based on no less than three approximations: (1) the binomial formula as an estimate for likelihood, (2) the binomial distribution as an estimate for the normal distribution, (3) the quadratic approximation as an estimate for the normal distribution. However, the approximations (1) and (3) provide exact rather than estimated likelihoods, and it turns out from the above examples that de log likelihood ratio test is, indeed, more sensitive than the standard tests. In addition,

the log transformation of the exponential binomial data is convenient, because exponents become simple multiplification factors. Also, the quadratic approximation is convenient, because an exponential equation is turned into a simpler quadratic equation (parabola).

Likelihood ratio statistics has a relatively short history. It was begun independently by Barnard[9] and Fisher[10] in the past World War II era. In this paper the log likelihood ratio test was used for the analysis of events only. The test can be generalized to other types of data including continuous data and the data in regression models, whereby the advantage of better sensitivity remains equally true. The test is, therefore, increasingly important in modern statistics.

We conclude that the log likelihood ratio test is more sensitive than traditional statistical tests including the t-(and z)-test, chi-square test and odds ratio test. Other advantages are the following: exponents can be conveniently handled by the log transformation and an exponential equation is turned into a simpler quadratic equation. A potential disadvantage of numerical problems is avoided by taking ratios of likelihoods instead of separate likelihoods in the final analysis.

7. CONCLUSIONS

Traditional statistical tests for the analysis of cardiovascular events have limited sensitivity, particularly with smaller samples. Exact tests, although infrequently used so far, should have better sensitivity, because they do not include standard errors as a measure of uncertainty. The log likelihood ratio test is one of them. The objective of the current chapter was to assess the above question using real and hypothesized data examples. In three studies of clinical events the log likelihood ratio test was consistently more sensitive than traditional tests, including the chi-square and the odds ratio test, producing p-values respectively between <0.05 and <0.002 and between not-significant and <0.05. This was true both with larger and smaller samples. Other advantages of the log likelihood ratio were: exponents can be conveniently handled by the log transformation and an exponential equation is turned into a simpler quadratic equation. A potential disadvantage of numerical problems is avoided by taking in the final analysis the ratios of likelihoods instead of separate likelihoods. Log likelihood ratio tests are consistently more sensitive than traditional statistical tests. We hope that the paper will stimulate cardiovascular researchers to more often apply them.

8. REFERENCES

1. BUGS y WinBUGS. http://www.mrc-bsu.cam.ac.uk/bugs http://cran.r-project.org
2. S-plus.http://www.mathsoft.com/splus
3. Stata. http://www.stata.com
4. StatsDirect. http://www.camcode.com
5. StatXact. http://www.cytel.com/products/statxact/statact1.html
6. True Epistat. http://ic.net/~biomware/biohp2te.htm

7. SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas
8. SPSS Statistical Software. http://www.spss.com
9. Barnard GA. A review of sequential analysis. J Am Stat Ass 1947; 422: 658-64.
10. Fisher RA. Statistical methods and scientific inferences. Ed by Oliver & Boyd, Edinburgh, UK, 1956.

CHAPTER 5

EQUIVALENCE TESTING

1. INTRODUCTION

A study unable to find a difference is not the same as an equivalent study. For example, a study of 3 subjects does not find a significant difference simply because the sample size is too small. Equivalence testing is particularly important for studying the treatment of diseases for which a placebo control would unethical. In the situation a new treatment must be compared with standard treatment. The latter comparison is at risk of finding little differences.

Figure 1 gives an example of a study where the mean result is little different from 0. Is the result equivalent then. H1 represent the distribution of our data and H0 is the null-hypothesis (this approach is more fully explained in chapter 2). What we observe is that the mean of our trial is only 0.9 standard errors of the mean (SEMs) distant from 0. which is far too little to reject the null-hypothesis. Our result is not significantly different from 0. Whether our result is equivalent to 0, depends on our prior defined criterium of equivalence. In the figure D sets the defined interval of equivalence. If 95% CIs of our trial is completely within this interval, we conclude that equivalence is demonstrated. This mean that with D_1 boundaries we have no equivalence, with D_2 boundaries we do have equivalence. The striped area under curve = the socalled 95 % CIs = the interval approximately between - 2 SEMs and + 2 SEMs (i.e., 1.96 SEMs with normal distributions, a little bit more than 2 SEMs with t-distributions. It is often hard to prior define the D boundaries, but they should be based not on mathematical but rather on clinical arguments, i.e., the boundaries where differences are undisputedly clinically irrelevant.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 73–79.

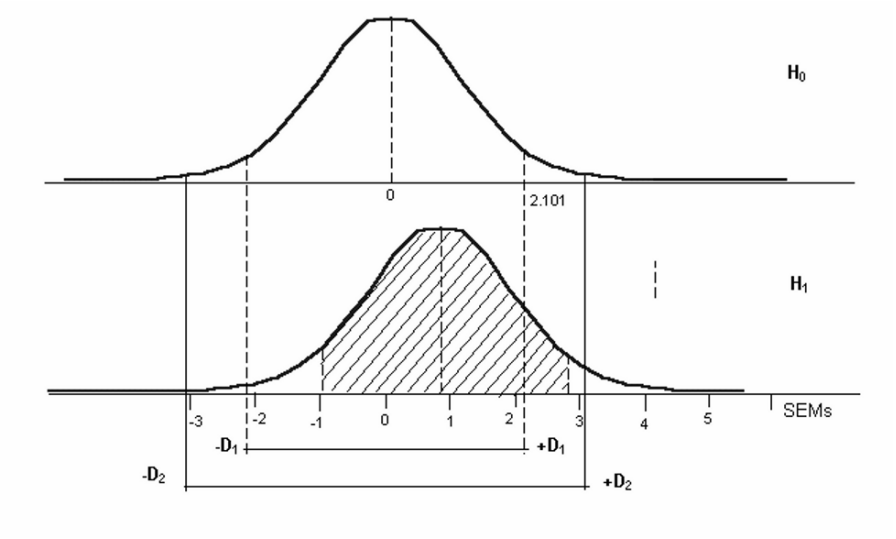


Figure 1. Null-hypothesis testing and equivalence testing of a sample of t-distributed data.

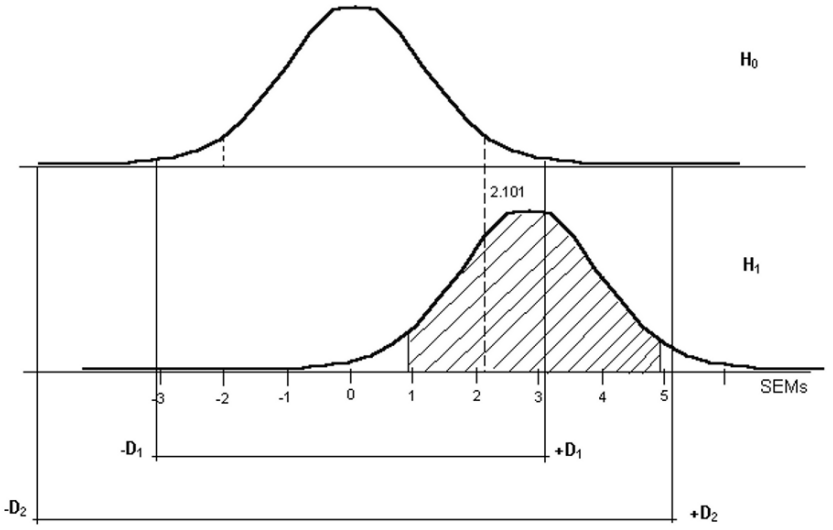


Figure 2. Null-hypothesis testing and equivalence testing of a sample of t-distributed data.

Figure 2 gives another example. The mean result of our trial is larger now: mean value is 2.9 SEMs distant from 0, and, so, we conclude that the difference from 0 is > approximately 2 SEMs and, that we can reject the null-hypothesis of no difference. Does this mean that our study is not equivalent? This again depends on our prior defined criterium of equivalence. With D_1 the trial is not completely within the boundaries and equivalence is thus not demonstrated. With D_2 the striped area of the trial is completely within the boundaries and we conclude that equivalence has been demonstrated. Note that with D_1 we have both significant difference and equivalence.

2. OVERVIEW OF POSSIBILITIES WITH EQUIVALENCE TESTING

Table 1 shows that any confidence interval (95 % CIs intervals between the brackets in each of the examples) that does not overlap zero is statistically different from zero. Only intervals between the prespecified range of equivalence –D to + D present equivalence. Thus, situations 3, 4 and 5 demonstrate equivalence, while 1 and 2, just like 6 and 7 do not. Situations 3 and 5 present equivalence and at the same time significant difference. Situation 8 presents nor significant difference, nor equivalence.

Table 1. Any confidence interval (95 % CIs intervals between the brackets in each of the examples) that does not overlap zero is statistically different from zero. Only intervals between the prespecified range of equivalence –D to + D present equivalence

Study (1-8)	Statistical significance demonstrated	equivalence demonstrated
1.	Yes	< not equivalent >
2.	Yes	< uncertain >
3.	Yes	< equivalent >
4.	No	< equivalent >
5.	Yes	< equivalent >
6.	Yes	< uncertain >
7.	Yes	< not equivalent >
8.	No	< uncertain >

-D O +D

true difference

Testing equivalence of two treatments is different from testing their difference. We will in this chapter use the term comparative studies to name the latter kind of studies. In a comparative study we use statistical significance tests to determine whether the null hypothesis of no treatment difference can be rejected, frequently

together with 95% CIs to better visualize the size of the difference. In an equivalence study this significance test has little relevance: failure to detect a difference does not imply equivalence; the study may have been too small with corresponding wide standard errors to allow for such a conclusion. Also, not only difference but also equivalence are terms that should be interpreted within the context of clinical relevance. For that purpose we have to predefine a range of equivalence as an interval from –D to +D. We can then simply check whether our 95% CIs as centered on the observed difference lies entirely between –D and +D. If it does equivalence is demonstrated if not, there is room for uncertainty. The above table shows the discrepancies between significance and equivalence testing. The procedure of checking whether the 95% CIs are within a range of equivalence does look somewhat similar to a significance testing procedure, but one in which the role of the usual null and alternative hypothesis are reversed. In equivalence testing the relevant null hypothesis is that a difference of at least D exists, and the analysis is targeted at rejecting this "null-hypothesis". The choice of D is difficult, is often chosen on clinical arguments: the new agent should be sufficiently similar to the standard agent to be clinically indistinguishable.

3. CALCULATIONS

95% CIs intervals are calculated according to the standard formulas

Continuous data paired or unpaired and normal distributions (with t-distribution **2**, which is actually 1.96, should be replaced by the appropriate t-value dependent upon sample size).

$$\text{Mean}_1\text{- mean}_2 \pm \mathbf{2} \text{ SEMs where}$$

$$\text{SEM}_{\text{unpaired differences}} = \sqrt{\text{SD}_1^2/n_1 + \text{SD}_2^2/n_2}$$

$$\text{SEM}_{\text{paired differences}} = \sqrt{\frac{(\text{SD}_1^2 + \text{SD}_2^2 - 2r \cdot \text{SD}_1 \cdot \text{SD}_2)}{n}} \quad \text{if } n_1=n_2=n$$

Binary data

$$\text{SEM}_{\text{of differences}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\text{With 95\% CIs :} \quad p_1 - p_2 \pm 2.\ \text{SEM}$$

More details about the calculation of SEMS of samples are given in chapter 1.

The calculation of required samples size of the trial based on expected treatment effects in order to test our hypothesis reliably, will be explained in the next chapter together with sample size calculations for comparative studies.
It is helpful to present the results of an equivalence study in the form of a graph (Table 1). The result may be:

1. The confidence interval for the difference between the two treatments lies entirely between the equivalence range so that we conclude that equivalence is demonstrated.
2. The confidence interval covers at least several points outside the equivalence range so that we conclude that a clinically important difference remains a possibility, and equivalence cannot be safely concluded.
3. The confidence interval is entirely outside the equivalence range.

4. EQUIVALENCE TESTING, A NEW GOLD STANDARD?

The classic gold standard in drug research is the randomized placebo controlled clinical trial. This design is favored for confirmatory trials as part of the phase III development of new medicines. Because of the large numbers and classes of medicines already available, however, new medicines are increasingly being developed for indications for which a placebo control group would be unethical. In such situations an obvious solution is to use as comparator an existing drug already licensed and regularly used for the indications in question. When an active comparator is used, the expectation may sometimes be that the new treatment will be better than the standard, the objective of the study may be to demonstrate this. This situation would be similar to a placebo control and requires no special methodology. More probably, however, the new treatment is expected to simply largely match the efficacy of the standard treatment but to have some advantages in terms of safety, adverse effects, costs, pharmacokinetic properties. Under these circumstances the objective of the trial is to show equivalent efficacy.

5. VALIDITY OF EQUIVALENCE TRIALS

A comparative trial is valid when it is blinded, randomized, explicit, accurate statistically and ethically. The same is true for equivalence trial. However, a problem arises with the intention to treat analysis. Intention to treat patients are analyzed according to their randomized treatment irrespective of whether they actually received the treatment. The argument is that it mirrors what will happen when a treatment is used in practice. In a comparative parallel group study the inclusion of protocol violators in the analysis tend to make the results of the two treatments more similar. In an equivalence study this effect may bias the study towards a positive result, being the demonstration of equivalence. A possibility is to carry out both intention-to-treat-analysis and completed-protocol-analysis. If no difference is demonstrated, we conclude that the study's data are robust (otherwise called sensitive, otherwise called precise), and that the protocol-analysis did not

introduce major sloppiness into the data. Sometimes, efficacy and safety endpoints are analyzed differently: the former according to the protocol analysis simply because important endpoint variables are missing in the population that leaves the study early, and intention to treat analysis for the latter, because safety variables frequently include items such as side effects, drop-offs, morbidity and mortality during trial. Either endpoint can of course be assessed in an equivalence assessment trial, but we must consider that an intention to treat analysis may bias the equivalence principle towards overestimation of the chance of equivalence.
Note: statistical power of equivalence testing is explained in the next chapter.

6. SPECIAL POINT: LEVEL OF CORRELATION IN PAIRED EQUIVALENCE STUDIES

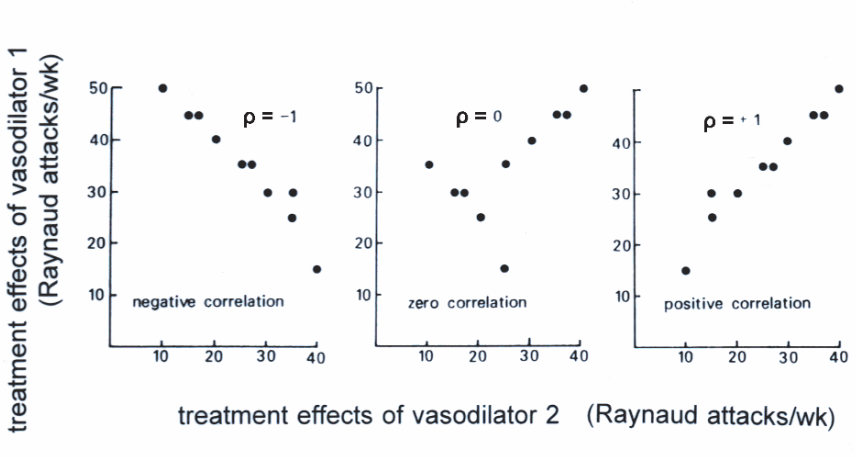


Figure 3. Example of 3 crossover studies of two treatments in patients with Raynaud's phenomenon. The Pearson's correlation coefficient ρ varies from -1 to 1.

Figure 3 shows the results of three crossover trials with two drugs in patients with Raynaud's phenomenon. In the left trial a negative correlation exists between the treatments, in the middle trial the correlation level is zero, while in the right trial a strong postive correlation is observed. It is calculated that the mean difference between the treatments in each trial equals 5 Raynaud attacks/week but that the standard errors of the differences are different, left trial 6.46, middle trial 2.78, right trial 0.76 Raynaud attacks / week. Figure 4 shows that with a D-boundary of ± 10 Raynaud attacks / week only the positive correlation study is able to demonstrate equivalence. Fortunately, most crossover studies have a positive

correlation between the treatments, and, so, the crossover design is generally quite sensitive to assess equivalence.

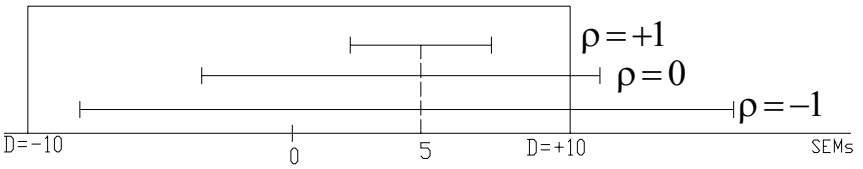


Figure 4. The mean difference between the two treatments of each of the treatment comparison of Figure 3 is 5 Raynaud attacks / week. However, standard errors, and, thus, 95% confidence intervals are largely different. With a D-boundary of ± 10 Raynaud attacks /week only the positive correlation study (ρ = +1) can demonstrate equivalence.

7. CONCLUSIONS

1. The use of placebos is unethical if an effective active comparator is available.
2. With an active comparator the new treatment may simply match the standard treatment.
3. Predefined areas of equivalence have to be based on clinical arguments.
4. Equivalence testing is indispensable in drug development (for comparison versus an active comparator).
5. Equivalence trials have to be larger than comparative trials. You will understand this after reviewing the next chapter.

CHAPTER 6

STATISTICAL POWER AND SAMPLE SIZE

1. WHAT IS STATISTICAL POWER

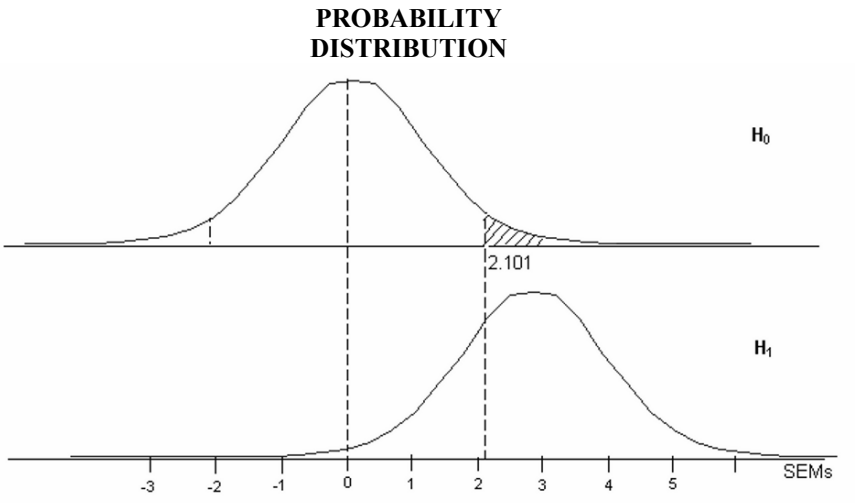


Figure 1. H1 is the given distribution of our data with mean value of 2.901 (= t = mean/SEM). β= area under curve (AUC) of H1 left from the dotted vertical line= ± 0.3 (± 30% of the total AUC). 1-β= ± 0.7=± 70% of total AUC of H1. Statistical power = ± 0.7 = chance of finding a difference when there is one.

Figure 1 shows 2 graphs of t-distributions. The lower graph (H1) could be a probability distribution of a sample of data or of a sample of paired differences between two observations. N=20 and so 95% of the observations is within 2.901 ± 2.101 standard errors of the mean (SEMs) on the x-axis (usually called z-axis in statistics). The upper graph is identical, but centers around 0 instead of 2.901. It is called the null-hypothesis H0, and represents the data of our sample if the mean results were not different from zero. However, our mean result is 2.901 SEMs distant from zero. If we had many samples obtained by similar trials under the same null-hypothesis, the chance of finding a mean value of more than 2.101 is < 5%, because the area under the curve (AUC) of H0 right from 2.101 <5% of total AUC. We, therefore, reject the assumption that our results indicate a difference just by chance and decide that we have demonstrated a true difference. What is the

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 81–97.

power of this test. The power has as prior assumption that there is a difference from zero in our data. What is the chance of demonstrating a difference if there is one. If our experiment would be performed many times, the distribution of obtained mean values of those many experiments would center around 2.901, and about 70% of the AUC of H1 would be larger than 2.101. When smaller than 2.101, our statistical analysis would not be able to reject the null-hypothesis of no difference, when larger, it would rightly be able to reject the null-hypothesis of no difference. So, in fact 100-70=30% of the many trials would erroneously be unable to reject the null-hypothesis of no difference, even when a true difference is in the data. We say the power of this experiment= 1-0.3=0.7 (70%), otherwise called the chance of finding a difference when there is one (area under curve (1-β)x 100%). β is also called the chance of making a type II error = chance of finding no difference when there is one. Another chance is the chance of finding a difference where there is none, otherwise called the type I error (area under the curve (2x α/2)x 100%). This type of error is usually set to be 0.05 (5%).

2. EMPHASIS ON STATISTICAL POWER RATHER THAN NULL-HYPOTHESIS TESTING

Generally, statistical tests reach their conclusions by seeing how compatible the observations were with the null-hypothesis of no treatment effect or treatment difference between test-treatment and reference-treatment. In any test we reject the null-hypothesis of no treatment effect if the value of the test statistic (F, t, q, or chi-square) was bigger than 95% of the values that would occur if the treatment had no effect. When this is so, it is common for medical investigators to report a statistically significant effect at P (probability) <0.05 which means that the chance of finding no difference if there is one, is less than 5%. On the other hand, when the test statistic is not big enough to reject this null-hypothesis of no treatment effect, the investigators often report no statistically significant difference and discuss their results in terms of documented proof that the treatment had no effect. All they really did, was fail to demonstrate that it did have an effect. The distinction between positively demonstrating that a treatment had no effect and failing to demonstrate that it does have an effect, is subtle but very important, especially with respect to the small numbers of subjects usually enrolled in a trial. A study of treatments that involves only a few subjects and then fails to reject the null hypothesis of no treatment effect, may arrive at this result because the statistical procedure lacked power to detect the effect because of a too small sample size, even though the treatment did have an effect.

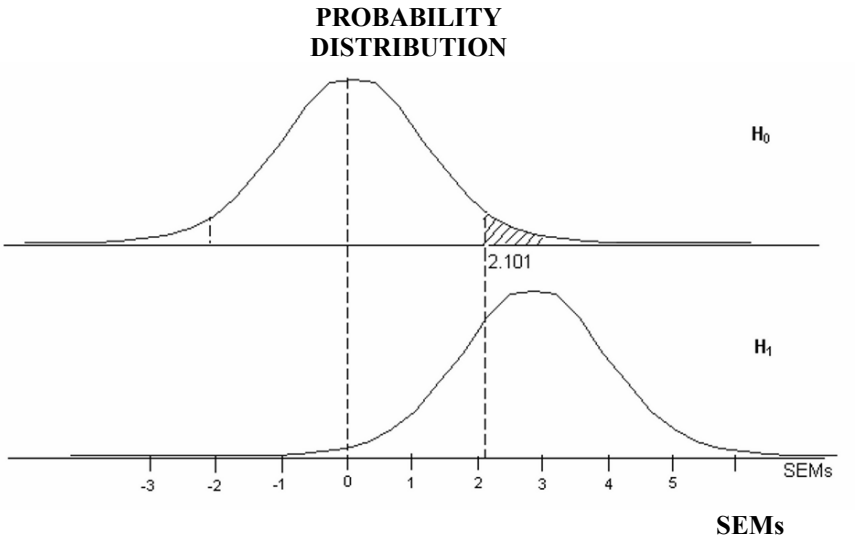


Figure 2. Example of t-distribution with n=20 and its null-hypothesis of no effect. Lower curve H1 or actual SEM distribution of the data, upper curve H0 or null-hypothesis of the study.

Figure 2 gives an example of a t-distribution with n=20 (H1) and its null-hypothesis of no effect (H0). 95% of all similar trials with no significant treatment difference from zero must have their means between –2.101 and +2.101 SEMs from zero. The chance of finding a mean value of 2.101 SEMs or more is 5% or less (α= 0.05 or α. 100%= 5%, where α is the chance of finding a difference when there is none = erroneously rejecting the null-hypothesis of no effect, also called type I error). The figure shows that in this particular situation the chance of β is 0.5 or β times 100% = 50% (β is the chance of finding no difference where there is one = the chance of erroneously accepting the null-hypothesis of no treatment difference, also called type II error).

Statistical power, defined as 1-β, can be best described as the chance of finding a difference where there is one = the chance of rightly rejecting the null-hypothesis of no effect. The figure shows that this chance of detecting a true-positive effect, i.e., reporting a statistically significant difference when the treatment really produces an effect is only 50%, and likewise that the chance of no statistically significant difference is no less than 50% either (β=0.5). It means that if we reject the null-hypothesis of no effect at P=0.05, we still have a chance of 50% that a real effect in our data is not detected. As a real effect in the data rather than no effect is the main underlying hypothesis of comparative drug trials, a 50% chance to detect it, is hardly acceptable for reliable testing. A more adequate cut-off level of rejecting would be, e.g., a 90-95% power level, with corresponding α level of

0.005 to 0.001. Many physicians and even some investigators never confront these problems because they never heard of power. An additional advantage of power analysis is the possibility to use power computations on hypothesized results a priori in order to decide in advance on sample size for a study.

3. POWER COMPUTATIONS

Calculating power can be best left over to a computer, because other approaches are rather imprecise. E.g., with normal distributions or t-distributions power =1-β can be readily visualized from a graph as estimated percentage of the (1-β) x 100% area under the curve. However, errors as large as 10-20 % are unavoidable with this approach. We may alternatively use tables for t- and z-distributions, but as tables give discrete values this procedure is rather inaccurate either.
A computer will make use of the following equations.

For t-distributions of continuous data

Power = 1-β = 1- probability [$z_{power} \leq (t\text{-}t^1)$] = probability [$z_{power} > (t\text{-}t^1)$]

where z_{power} represents a position on the x-axis of the z-distribution (or in this particular situation more correctly t-distribution), and t^1 represents the level of t that for the given degrees of freedom (≈ sample size) yields an α of 0.05. Finally, t in the equation is the actual t as calculated from the data.
Let's assume we have a parallel-group data comparison with test statistic of t=3.99 and n=20 (P<0.001). What is the power of this test? $Z_{power} = (t^1\text{-}t) = 3.99\text{-}2.101 = 1.89$. This is so, because t^1= the t that with 18 degrees of freedom (dfs) (n= 20, 20-2) yields an α of 0.05. To convert z_{power} into power we look up in the t-table with dfs=18 the closest level of probability and find approximately 0.9 for 1.729. The power of this test thus is approximately 90%.

For proportions

$$z_{power} = 2.\ (\text{arcsine}\ \sqrt{p_1}\ -\ \text{arcsine}\ \sqrt{p_2}\ )\ \sqrt{\frac{n}{2}}\ -\ z^1$$

where z_{power} is a position on the x-axis of the z-distribution and z^1 is 2 if α=0.05 (actually 1.96). It is surprising that arcsine (= 1/sine) expressed in radians shows up but it turns out that power is a function of the square roots of the proportions, which has a 1/sine like function.
A computer turns z_{power} into power. Actually, power graphs as presented in many current texts on statistics can give acceptable estimates for proportions as well.

For equivalence testing of samples with t-distributions and continuous data

Power= 1-β= 1- probability [z< (D/SEM - $z_{1-\alpha}$)]

where z is again a position on the x-axis of the z- or t-distribution, D is half the interval of equivalence (see previous chapter), and $z_{1-\alpha}$ is 2 (actually 1.96) if α is set at 5%.

4. EXAMPLES OF POWER COMPUTATION USING THE T-TABLE

FIRST EXAMPLE

Although a table gives discrete values, and is somewhat inaccurate to precisely calculate the power size, it is useful to master the method, because it is helpful to understand what statistical power really is. The example of Figure 3 is given. Our trial mean is 2.878 SEMs distant from 0 (= the t-value of our trial). We will try to find beta by subtracting t - t^1 where is the t - value that yields an area under the curve (AUC) of 5% = 2.101. t - t^1 = 2.878-2.101=0.668. Now we can use the t-table to find 1-beta = power.

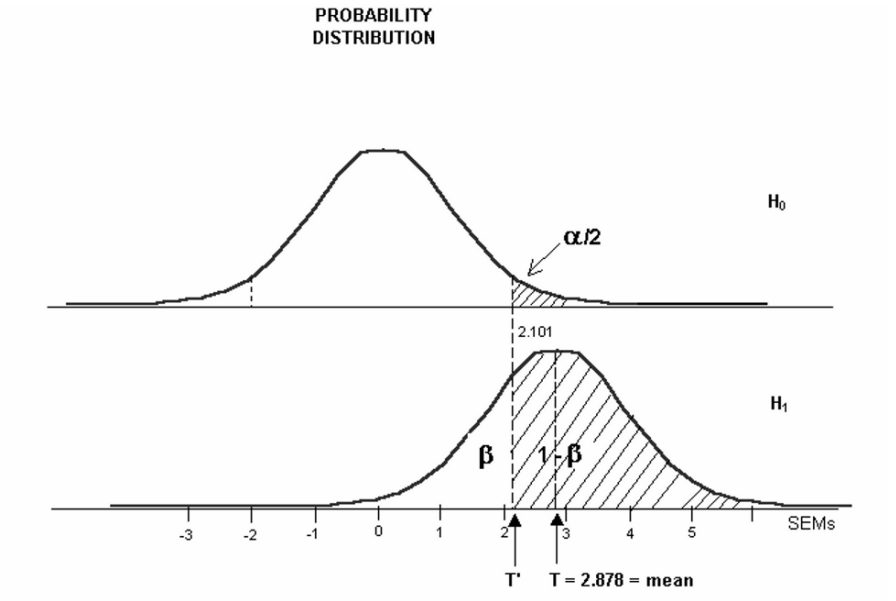


Figure 3. Example of power computation using the t-table.

The t-table (Table 1) gives 8 columns of t - values and one column (left one) of degrees of freedom. The upper rows give an overview of AUCs corresponding to various t-values and degrees of freedom. In our case we have two groups of 10 subjects and thus 20-2=18 degrees of freedom (dfs). The AUC right from 2.101 =

0.05 (tested 2-sided = tested for both > + 2.101 and < - 2.101 distant from 0). Now for the power analysis. The t - value of our trial = 2.878. The t^1 – value = approximately 2.101; t - t^1 = approximately 0.777. The AUC right from 0.777 is right from 0.688 corresponding with an area under the curve (AUC) < 0.25 (25%). Beta, always tested one-sided, is, thus, <25% ; 1 - beta = power = > 75%.

Table 1. t-table: v= degrees of freedom for t-variable, Q=proportion of cases cut off on the upper tail of the t- distribution

	Q = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	2Q = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

SECOND EXAMPLE

The mean result from the example in Figure 4 is 2.1 SEMs distant from zero, which is equal to the t - value in the underneath t-table. We find beta by subtracting $t - t^1$ where t^1 is the t yielding AUC of 5% =2.101. $t - t^1 = 0.0$. Now we use t - table to find 1 - beta. The t – va lue = 2.1, $t^1 = 2.1$, $t - t^1 = 0.0$, close to 0.257.

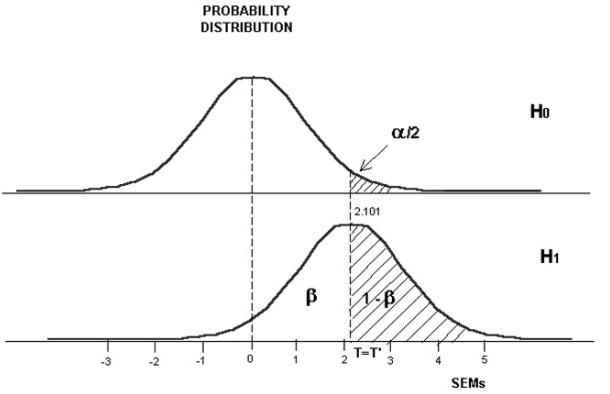


Figure 4. Example of power computation using the t-table.

The AUC is, thus, close to 0.4 and will be approximately 0.50. Beta (1-sided) = approximately 50%, 1-beta = power = 1- 0.50 = approximately 0.50 = approximately 50%, power is 50%. This little power is not acceptable for accurate testing.

Table 2. t-table: v= degrees of freedom for t-variable, Q=proportion cases cut off on the upper tail of the t- distribution

ν = 0.5	Q = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	2Q = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.0 .257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

THIRD EXAMPLE

Things may get worse. The mean result of the study from Figure 5 is 0.9 SEMs distant from zero. The t-value = 0.9. We find beta by subtracting t - t^1 where t^1 is the t yielding an AUC of 0.05 = 2.101; t - t^1 = -1.20.

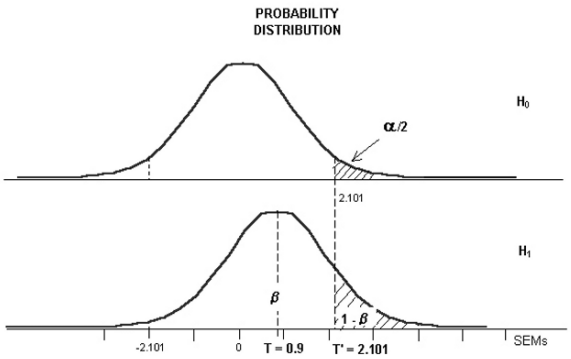


Figure 5. Example of power computation using the t-table.

Our t - value is, thus, 0.9 , t^1 is 2.1, t - t^1 = -1.2, 1.2 is between 0.68 and 1.33, and close to 1.33, and corresponds with an AUC a bit more than 10%: 15% or so, -1.2 corresponds with an AUC 100% - 15% = 85%, beta = 85% , 1 - beta = 15% = STATISTICAL POWER. Notice that this procedure is getting rather imprecise with extreme values.

Table 3. t-table: v= degrees of freedom for t-variable, Q=proportion of cases cut off on the upper tail of the t- distribution

	Q = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.001
	2Q = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

5. CALCULATION OF REQUIRED SAMPLE SIZE, RATIONALE

An essential part of planning a clinical trial is to decide how many people need to be studied in order to answer the study objectives. Just pulling the sample sizes out of a hat gives rise to:

1. Ethical problems, because if too many patients are given a potentially inferior treatment, this is not ethical to do.
2. Scientific problems, because negative studies require the repetition of the research.
3. Financial problems, because extra costs are involved in too small and too large studies.

If we have no prior arguments to predict the outcome of a trial, we at least will have an idea of the kind of result that would be clinically relevant. This is also a very good basis to place prior sample size requirement on. E.g., a smaller study, for example, will be needed to detect a fourfold increase than a twofold one. So the sample size also depends on the size of result we want to demonstrate reliably.

6. CALCULATIONS OF REQUIRED SAMPLE SIZE, METHODS

An essential part of planning a clinical trial is to decide: how many people need to be studied in order to answer the study objectives.

A simple method:
Mean should be at least 1.96 or approximately 2 SEMs distant from 0 to obtain statistical significance.
Assume: mean = 2 SEM
Then mean/ SEM=2
Then mean/ SD/ $\sqrt{n} = 2$
Then $\sqrt{n}$= 2.SD/mean
Then **$n = 4.(SD/mean)^2$**
For example, with mean=10 and SD=20 we will need a sample size of at least n= 4 $(20/10)^2$ = 4 x 4 = **16**. P-value is then 0.05 but power is only 50%.

A more accurate method is the power index method:
The statistical power (1) of a trial assessing a new treatment versus control is determined by 3 major variables:
(2) D (mean difference or mean result),
(3) Variance in the data estimated as SD or SEM,
(4) Sample size.
It follows that we can calculate (4) if we know the other 3 variables.

The relationship between (4) and the 3 other variables can be expressed in fancy formulas with $(z_\alpha + z_\beta)^2$ = **power index** as an important element in all of them. Here is the formula for continuous variables

$$\mathbf{n = (SD/mean)^2 (z_\alpha + z_\beta)^2}$$

If the power index for null-hypothesis is $(z_\alpha + z_\beta)^2$, what is the size of this $(z_\alpha + z_\beta)^2$ **?**

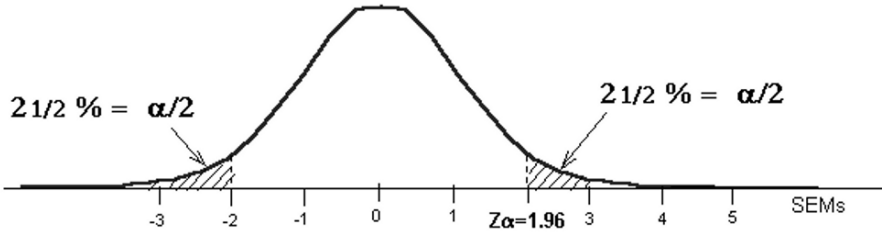


Figure 6. Calculating power indexes.

What does for example $Z_{(alpha)}$ exactly mean? $\mathbf{Z_{(alpha)}}$ **means " a place" on the Z-line. What place?** If **alpha** is defined 5%, or rather 2x2 1/2 % , then right from this place on the Z-line AUC = 5%, or rather 2x2 1/2 %. So this place must be 1.96 SEMs distant from 0, or a bit more with t-distribution. **So** $\mathbf{Z_{alpha}}$ **= 1.96= approximately 2.0** (Figure 6).

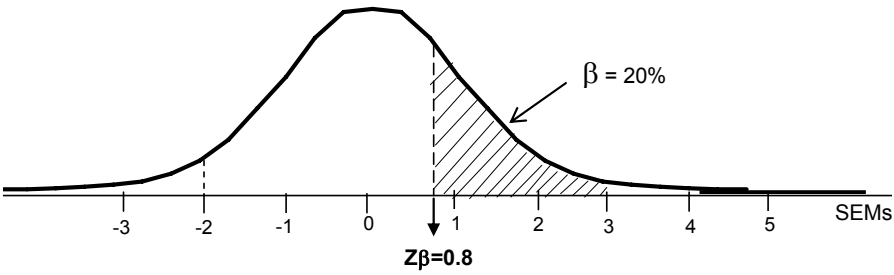


Figure 7. Calculating power indexes.

What does $Z_{(beta)}$ exactly mean? If **beta** is defined 20%, what is the place on Z-line of $Z_{(beta)}$? Right from this place the AUC= 20% of the total AUC. This means that this place must be approximately 0.6 SEMs distant from 0. **So** $\mathbf{Z_{beta}}$ **= approximately 0.8** (Figure 7).

Now we can calculate the power index $(z_\alpha + z_\beta)^2$.

$Z_{(alpha)}$ = approximately 2.0
$Z_{(beta)}$ = approximately 0.8
power index = $(z_\alpha + z_\beta)^2 = 2.8^2 = 7.8$
As the formula for continuous variables is $n = (SD/mean)^2 (z_\alpha + z_\beta)^2$,
we can now conclude that with $\alpha = 5\%$ and power = $1 - \beta = 80\%$ the required sample size is $n = 7.8\ (SD/mean)^2$. E.g., with SD=20 and mean=10, we will need a sample size of $n = 7.8\ (20/10)^2 =$ **32.**
So, accounting a power of 80% requires 32, rather than the 16 patients, required according to the simple method.

Power calculation for parallel-group studies:
For parallel-group studies including two groups larger sample sizes are required. Each group produces its own mean and standard deviation (SD).
The pooled $SD = \sqrt{(SD_{group1}^2 + SD_{group2}^2)}$
The equation for sample size is given by:

$$n = 2\ (z_\alpha + z_\beta)^2\ (\text{pooled SD} / \text{mean difference})^2$$

If the mean difference = 10, and the pooled $SD = \sqrt{(20^2 + 20^2)} = 28.3$, then the required sample size is given by

$$n = 2 \times 7.8 \times (28.3/10)^2 = 2 \times 7.8 \times 8.01 = 126$$

Thus, 63 subjects per group are required for the purpose of 80 % power with alpha = 0.05.

Required sample size equation for studies with proportions:
If we have arguments to expect events in 10% of the subjects included, then p = proportion = 0.1. The SD of this proportion = $\sqrt{[p(1-p)]}$.

The equations for continuous data and proportions are very similar.

Continuous data: $n = \text{powerindex} \times (SD / mean)^2$
Proportions : $n = \text{powerindex} \times (SD / proportion)^2$

So, if p = 0.10, then the required sample size is given by

$$n = 7.8 \times [(0.1 \times 0.9) / 0.1]^2$$

For parallel-group studies with two proportions we again have to pool the SDs.

pooled SDs = $\sqrt{\ (\ SD_{group1}^2 + SD_{group2}^2\ )}$

For example

	number of subjects with an event			
	yes	no		
group 1	a	b	proportion $p_1 = a / (a+b)$	$SD_1 = \sqrt{[p_1 (1-p_2)]}$
group 2	c	d	proportion $p_2 = c / (c+d)$	$SD_2 =$ .....
				pooled SD = $\sqrt{\ (\ SD_1^2 + SD_2^2\ )}$

It is hard to recognize the equation from the equation of continuous data, but it is actually very similar:

$$n = 2\,(z_\alpha + z_\beta)^2 \cdot \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2}$$

(where p_1 and p_2 are the proportions to be compared).

As an example a standard and new treatment are compared.
The standard treatment produces a proportion of responders of $p_1 = 0.1$, the new treatment of $p_2 = 0.2$. The required sample size is calculated according to

$$n = 2 \times 7.8 \times \frac{0.1\,(1-0.1) + 0.2\,(1-0.2)}{(0.1 - 0.2)^2} = 390$$

The required sample per group is, thus, 195.

Note that a requested power of 90 % means a power index of 10.5. In this study 526 subjects would have to be included.

Required sample size formula for equivalence testing:
N = 2 (between subject variance) $(z_{1-1/2\alpha} + z_{1-1/2\beta})^2 / D^2$
(where D is minimal difference we wish to detect).

What size is the **power index of equivalence test $(z_{1-1/2\alpha} + z_{1-1/2\beta})^2$?**
If the power index of equivalence testing = $(z_{1-1/2\alpha} + z_{1-1/2\beta})^2$
What is the size of this power index?

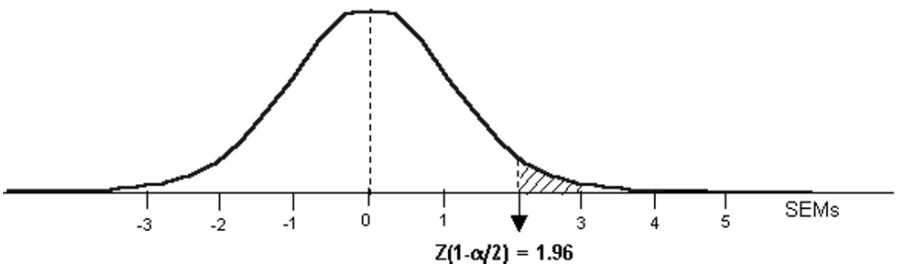


Figure 8. Calculating power indexes.

If alpha is defined 5%, then ½ alpha = 2 ½ %. What is the place on the Z-line of **$Z_{(1-1/2\alpha)}$? Left from this place the AUC = 1- ½ alpha = 100- 2 ½ %= 97 ½ % of total AUC.** So this place is, just like Z_{alpha} , 1.96 SEMs distant from 0, or bit more with t-distribution. So, **$Z_{(1-½\ alpha)}$ = 1.96 or approximately 2.0** (Figure 8).

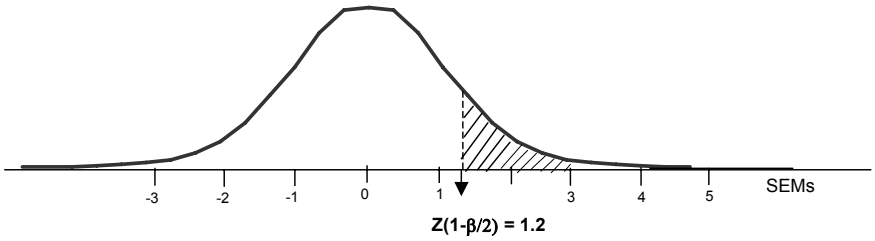


Figure 9. Calculating power indexes.

Now, if beta is defined 20%, then ½ beta = 10% What is the place on the Z-line of $Z_{(1-1/2\ beta)}$? **Left from the place the AUC = 100% -10% = 90% of total AUC.** This means that this place must be approximately 1.2 SEMs distant from 0, or a bit more, and, thus, **$Z_{(1-½\ beta)}$ = approximately 1.2** (Figure 9).
Now we can calculate this power index. $Z_{(1-½\ \alpha)}$ = approximately 2.0. $Z_{(1-½\ \beta)}$ = app 1.2. The power index for equivalence testing = $(2.0 + 1.2)^2$ = approximately 10.9.
NOTE: power index null hypothesis testing = 7.8
" " equivalence testing = 10.9

Obviously, for equivalence testing larger sample sizes are required !
Equivalence trials often include too few patients. The conclusion of equivalence becomes meaningless if, due to this, the design lacks power. Testing equivalence usually requires a sample larger than that of comparative null hypothesis testing studies. Required numbers of patients to be included should be estimated at the design stage of such studies.

7. TESTING INFERIORITY OF A NEW TREATMENT (THE TYPE III ERROR)

An inferior treatment may sometimes mistakenly be believed to be superior. "Negative"studies, defined as studies that do not confirm their prior hypotheses, may be "negative" because an inferior treatment is mistakingly believed to be superior. However, from a statistical point of view this possibility is unlikely, because the possibility of a type III error can not be rejected. Suppose in a study the mean results is + 1 SEM distant from the mean of the null hypothesis of no treatment effect (Figure 10).

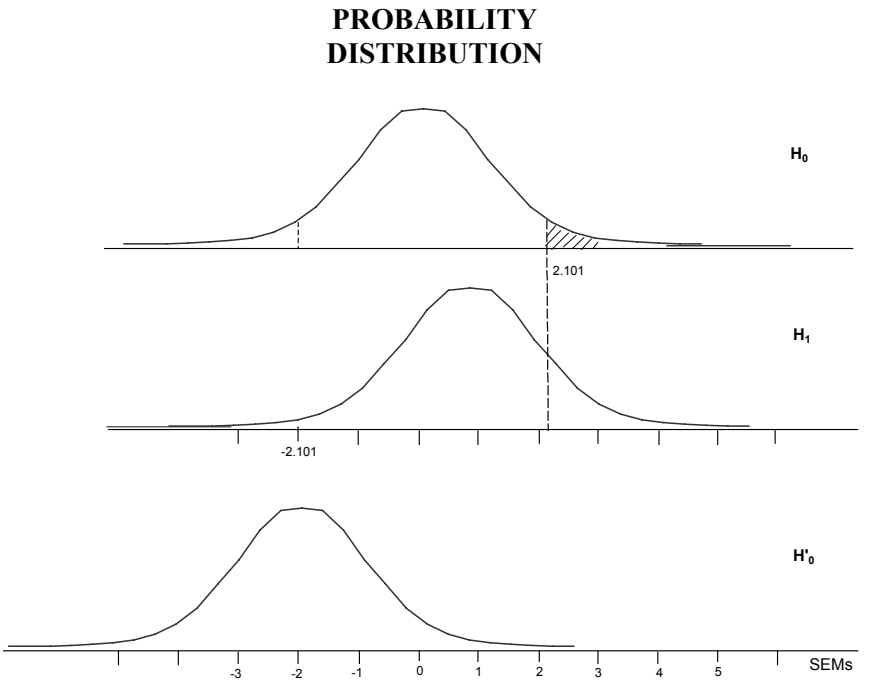


Figure 10. Study with n=20 and mean results + 1 SEM distant from the mean of the null-hypothesis of no treatment effect (H_0). For testing the chance that our treatment is significantly inferior, a new null hypothesis at approximately – 2SEMs left from zero is required.

This means that we are unable to reject this null hypothesis, because a null hypothesis is rejected when the mean result of a study is more than about 2 SEMs distant from zero ($P<0.05$), and the study is thus "negative". For testing the chance that our treatment is significantly inferior, a new null-hypothesis at approximately – 2SEMs distant from zero is required (Figure 10). This null-hypothesis is about 3 SEMs distant from our mean result, which means that this chance is <0.001. So, it seems that even statistically "negative" trials give strong evidence that the favored treatment is, indeed, not inferior. This issue can be illustrated by an example. The treatment of hypertension is believed to follow a J-shape curve, where overtreatment produces increased rather than reduced mortality/morbidity. A different theory would tell you that the more intensive the therapy the better the result. This latter theory was recently tested in the HOT trial[1] (HOT investigators, Lancet 1998; 87: 133), but could not be confirmed: high dosage antihypertensive therapy was not significantly better than medium-dosage therapy. Probably it was not worse either, however, unfortunately, this was not tested in the report. The

study would definitely have been powerful to test this question, and, moreover, it would have solved a major so far unsolved discussion.
An additional advantage of testing type III errors is, that it helps preventing well-designed studies from going down in history as just “negative” studies that did not prove anything and are more likely not to be published, leading to unnecessary and costly repetition of research. If such “negative” studies are capable of rejecting the chance of a type III error, they may be reconsidered as a study that is not completely negative and may be rightly given better priority for being published.

8. CONCLUSIONS

1. If underlying hypothesis is that one treatment is really different from control, power analysis is a more reliable to evaluate the data than null hypothesis testing; Power level of at least 80% is recommended. Power = chance of finding a difference where there actually is one.
2. Despite speculative character of prior estimates, it is inappropriate not to calculate required sample size based on expected results.
3. Type III error demonstrates in negative trial whether the new treatment is worse than control.
4. Important formulas:

Power = 1- prob ($z < (t - t^1)$) where prob = probability

Power index need for calculating sample size $(z_\alpha + z_\beta)^2$ is generally 7.8.

Required sample size = $2. (SD/mean)^2 (z_\alpha + z_\beta)^2$

5. Required knowledge after studying this chapter: to calculate power from simple example of (continuous) trial data using t-table, to calculate required sample size for continuous data trial with alpha=0.05 and beta = 0.20 using power index.

9. REFERENCES

1. HOT investigators. The HOT trial. Lancet 1998; 87: 133-142.

CHAPTER 7

INTERIM ANALYSES

1. INTRODUCTION

Clinical trials tend to have a long duration, because mostly patients are enrolled one by one, and their responses to treatment are observed sequentially. For the organizers this part of the trial is an exciting phase because after all the hard work involved in planning and getting the trial started, finally concrete data will become available. Immediately, there is the possibility to look at the data in order to check that the trial protocol is pursued appropriately by the investigators and to look at any difficulties, e.g., those with patient and/or doctor compliance, and to see whether there is any need for protocol alterations[1]. "Looking at the data" for such purposes should, however, be done carefully. In this chapter we will discuss questions such as:

1. why should we monitor a trial;
2. who should monitor a trial;
3. what should be monitored;
4. why should we be careful.

2. MONITORING

Careful conduct of a clinical trial according to the protocol has a major impact on the credibility of the results[2]; to ensure patient / doctor compliance with the protocol, careful monitoring of the trial is a prerequisite. In large-scale pharmaceutical phase III trials, mainly two types of monitoring are being used: one is concerned with quality assessment of trial, and the other with the assumptions that were made in the protocol concerning treatment differences, power, and adverse effects. The quality of the trial is greatly enhanced when checks are performed to ensure that

1. the protocol requirements are appropriately met by investigators and patients;
2. inclusion and exclusion criteria are appropriately met;
3. the rate of inclusion of patients in the trial is in accordance with the trial plan;
4. the data are being accrued properly, and;
5. design assumptions are met.

This type of monitoring does not require access to the data in the trial, nor is unblinding necessary, and therefore has no impact on the Type I error of finding a difference where there is none[2] (see also chapter 5.1, and 5.2 of the current book).

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 99–105.

Usually, this type of monitoring is carried out by a specialized monitoring team under the responsibility of the steering committee of the trial. The period for this type of monitoring starts with the selection of the trial centers and ends with the collection and cleaning of the last patient's data.
Inclusion and exclusion criteria should be kept constant, as specified in the protocol, throughout the period of patient recruitment. In very long-term trials accumulating medical knowledge either from outside the trial, or from interim analyses, may warrant a change in inclusion or exclusion criteria. Also, very low recruitment rates due to over-restrictive criteria, may sometimes favor some change in the criteria. These should be made without breaking the blinding of the trial and should always be described in a protocol amendment to be submitted to the ethic committee for their approval. This amendment should also cover any statistical consequences such as sample size, and alterations to the planned statistical analysis.
The rate of subject accrual should be monitored carefully, especially with long-term trials. If it falls below the expected level, the reasons why so should be identified, and action taken not to jeopardize the power of the trial. Naturally, the quality of the data should be assessed carefully. Attempts should be made to recover missing data and to check the consistency of the data.

3. INTERIM ANALYSIS

The other type of monitoring requires the comparison of treatment results, and it, therefore, generally requires at least partly unblinded access to treatment group assignment. This type of monitoring is actually called interim analysis. It refers to any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to formal completion of the trial.
The primary goals for monitoring trial data through interim analysis include

1. ethical concerns to avoid any patient receiving a treatment the very moment it is recognized to be inferior;
2. (cost-)efficiency concerns of avoiding undue prolongation of a trial once the treatment differences are reasonably clear-cut, and;
3. checking whether prior assumptions concerning sample size, treatment efficacy and adverse effects are still valid.

As the sample-size of the trial is generally based on preliminary and/or uncertain information, an interim check on the unblinded data may also be useful to reveal whether or not overall response variances, event rates or survival experience are as anticipated. A revised sample size may then be required using suitable modified assumptions. As a matter of course, such modification should be documented in a protocol amendment and in the clinical study report. Steps taken to preserve blindness during the rest of the trial and consequences for the risk of type I errors and the width of the confidence intervals should be accounted for.
Particularly, severe toxic reactions, as well as other adverse effects, are important and need careful observation and reporting to the steering committee, so that prompt

action can be taken. Investigators need to be warned to look out for such events and dose modifications may be necessary.
Every process of examining and analyzing data as accumulated in a clinical trial, either formally or informally, can introduce bias and/or increase of type I errors. Therefore, all interim analyses, formal or informal, preplanned or ad hoc, by any study participant, steering committee member, or data monitoring group should be described in full in the clinical study report, even if their results were disclosed to the investigators while on trial [3].
For the purpose of reducing the risk of biases there are a number of important points in the organisation of the analysis and the interpretation of its results to keep in mind.

I - In most trials there are many outcome variables, but in interim analyses it is best to limit the number to only the major variables in order to avoid the multiple comparison problem (referred to in chapter 1.1). Pocock[1] recommends to use only one main treatment comparison for which a formal 'stopping rule' may be defined, and to use the other treatment comparisons only as an informal check on the consistency of any apparent difference in the main comparison.

II - It is important to perform the interim analysis on correct and up-to-date data. The data monitoring and data checks should be performed on all of the data generated at the time of the interim analysis in order to avoid any selection bias in the patients.

III - The interim analysis should be performed only when there is a sufficient number of patients. Any comparison is academic when the sample size is so small that even huge treatment differences will not be significant.

IV - The interim analysis should not be too elaborate, because there is a limited goal, namely to check whether differences in the main treatment comparison are not huge to the extent that further continuation of the trial would seem unethical.

V - The interim analysis should be planned only when a decision to stop the trial is a serious possibility. With very long-term treatment periods in a trial when the period between patient entry and observance of patient outcome is very long, the patient accrual may be completed before any interim analysis can be performed and the interim analysis results will have no impact on the trial anymore.

VI - The decision to stop the trial must be made according to a predefined stopping rule. The rule should be formulated in terms of magnitude and statistical significance of treatment differences and must be considered in the light of adverse effects, current knowledge, and practical aspects such as ease of administration, acceptability and cost. We must decide in advance what evidence of a treatment difference is sufficiently strong to merit stopping the trial. Statistical significance is a commonly used criterion, but the usual P-level is not appropriate. The problem with

statistical significance testing of interim data is that the risk of a type I error may be considerably increased because we perform more than one analysis. Hence, for a sequence of interim analyses we must set a more stringent significance level than the usual P<0.05. We may use a Bonferroni adjustment (see also Chapter 1 introduction), i.e., use as significance level the value 0.05 divided by the number of planned interim analyses, but this leads in most cases to a somewhat overconservative significance level. Therefore, in most trials a so-called group-sequential design is employed. This subject will be discussed in the next section. A practical guideline is to use Pocock's criteria[4]: **if one anticipates no more than 10 interim analyses and there is one main response variable, one can adopt p < 0.01 as the criterion for stopping the trial.** An example of this approach is the following: "stop the trial if the treatment difference is 20% or larger and this difference is statistically significant with a p - value less than 0.01, and the proportion patients with adverse effects is less than 10%." The outcome of the interim analysis may also be such that the treatments differ far less than expected. In such case the trial might be stopped for lack of efficacy. Again, it is essential that a formal stopping rule is formulated in advance specifying the boundary for the treatment difference for the given confidence intervals (CIs). In this case statistical significance is not helpful as an additional criterion, but it is helpful to calculate the confidence interval of the observed treatment difference and to see whether the expected treatment difference, specified in the protocol, is far outside that interval.

VII - It is necessary to keep the results of the interim analysis as confidential as possible. Investigators may change their outlook and future participation to the trial, and might even change their attitudes towards treatment of patients in the trial if he/she is aware of any interim results. This may cause a serious bias to the overall trial results. The U.S. Food and Drug Administration (FDA) therefore recommends not only that the execution of the interim analysis be highly confidential[2], but also that the investigators not be informed about its results unless a decision to stop the trial has been made. An external independent group of investigators should ideally perform the interim analysis, for the benefit of the objectivity of the research (although complete independence may be an illusion, it is still better to have some other persons with their own ethical and scientific principles look at your data than do it yourself). The steering committee should be informed about the decisions to continue or discontinue or the implementation of protocol amendments only.

VIII - There is little advantage to be gained from carrying out a large number of interim analyses: the consequences of executing many interim analyses are that the sample sizes are small (at least in the first analyses), and that a smaller significance level must be used. Pocock[5] recommends never to plan more than 5 interim analyses, but at the same time to plan at least one interim analysis, in order to warrant scientific and ethical validity of the trial.

4. GROUP-SEQUENTIAL DESIGN OF INTERIM ANALYSIS

Group sequential design is the most widely used method to define the stopping rule precisely and it was introduced by Pocock.[7] The FDA[2] advocates the use of this design, though it is not the only acceptable type of design, and the FDA does so particularly for the purpose of safety assessment, one of its major concerns.

In a group-sequential trial we need to decide about the number (N) of interim analyses and the number (n) of patients per treatment that should be evaluated in between successive analyses: i.e. if the trial consists of two treatment arms 2n patients must be evaluated in each interim analysis. Pocock[7] (and extensively explained in Pocock[3] provides tables for the exact nominal significance levels depending on the number of interim analyses N and the overall significance level. For instance if a trial is evaluated using a normal distributed response variable with known variance and one wishes the overall significance level to be ,=0.05 and one plans N=2 analyses, then the nominal significance level must be set at 0.0294. If N=3 or 4 or 5, the nominal significance levels must be set at 0.0221, 0.0182, and 0.0158, respectively. For other types of response variables, Pocock[7] provides similar tables. Pocock[7] also provides tables of the optimal sample size numbers of patients to be included in successive interim analyses.

Several extensions of the practical rules of Pocock were developed, for instance rules for letting the nominal significance level vary between interim analyses. In practice a far more stringent p-value is suggested for earlier interim analyses and a less stringent one for later analyses. Pocock[1] claimed that such a variation might be sensible for studies with a low power, but that almost no efficiency is gained in studies with powers of 90% or higher. Other extensions concern one-sided testing[8] and skewed designs where a less stringent rule might be adopted for stopping if the new treatment is worse than the standard and a more stringent rule if the new treatment appears to be better than the standard.

5. CONTINUOUS SEQUENTIAL STATISTICAL TECHNIQUES

Historically, the statistical theory for stopping rules in clinical trials has been largely concerned with sequential designs for continuous monitoring of treatment differences. The basic principle is that after every additional patient on each treatment has been evaluated, some formal statistical rule is applied to the whole data so far to determine whether the trial should stop. The theory of sequential techniques is already quite old (developed in the early fourties and even earlier than that[9]), and many excellent textbooks have been published[10]; here we adopt the arguments of Whitehead.[11]

The central idea is to calculate after each additional patient (or after I additional patients) (a function of) the treatment difference, called Z, and the total amount of information, called V, sampled thus far. These two statistics are plotted graphically against each other each time a new patient is evaluated. The stopping rule of the trial entails evaluating whether a boundary is crossed. In Figure 1 a typical example of a sequential trial with a so-called triangular test is illustrated.

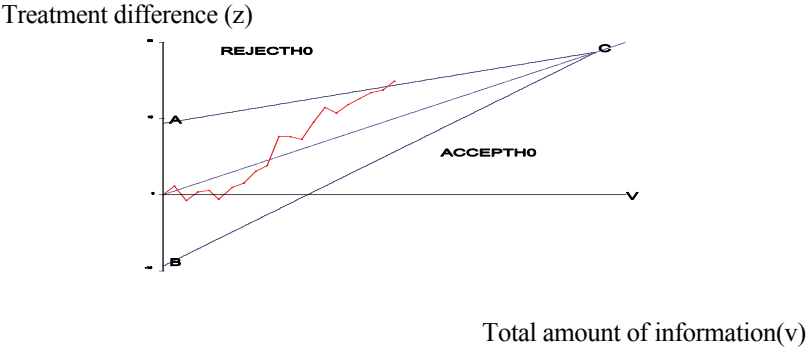


Figure 1. Typical example of a sequential trial with a so-called triangular test. The undulating line illustrates a possible realisation of a clinical trial: after each time a new patient could be evaluated, Z and V are calculated and the line is extended a little further. The line-sections AC and BC are the stopping boundaries, and the triangular region ABC is the continuation region. If the sample path crosses AC, the null hypothesis is rejected at the 5% significance level, and if BC is crossed then H_0 is accepted. When Z is replaced by t or chi-square statistic, and V by degrees of freedom, the graph represents very much the same as the t- or chi-square tables (appendix) respectively do.

The undulating line illustrates a possible realisation of a clinical trial: after each time a new patient could be evaluated, Z and V are calculated and the line is extended a little further. The line-sections AC and BC are the stopping boundaries, and the triangular region ABC is the continuation region. If the sample path crosses AC, the null hypothesis is rejected at the 5% significance level, and if BC is crossed then H_0 is accepted. The triangular test is one of many possible sequential trial designs; but the triangular test has some very attractive characteristics. If the treatment difference is large, it will lead to a steeply increasing sample path, and consequently to a small trial because the AC boundary is reached quickly. If there is no difference between treatment, the sample path will move horizontally and will cross the BC boundary quickly which also leads to a small trial. If the treatment difference is negative, the BC boundary will be crossed even quicker.

The trick is to devise sensible boundaries. Whitehead [11] gives an elaborate discussion on how to do this (as well as how to calculate Z and V). Whitehead [11] also discussed many different sequential plans for many different types of clinical trials and data-types. Whitehead and his associates have also developed a user-friendly computer program to design and analyze sequential clinical trials.[12]

6. CONCLUSIONS

Interim analyses in clinical trials can be of great importance in maintaining quality standards of the entire investigation and such analyses may be of crucial importance if clinical trials are to be ethically acceptable. Drawbacks of interim analyses are the increased risk of the type I error and the potential introduction of several kinds of biases, such as loss of validity factors, including blinding and randomization. It is rarely sensible to perform more than 5 interim analyses and usual 1 interim analysis before the final assessment suffices. It is crucial to specify in advance in the study protocol, how many analyses are to be performed and on how many patients, and which decisions are to be made on the basis of the interim results. It is best to let an external independent group, often called Independent Data Monitoring Committee (IDMC), execute the job and to keep its results as confident as is ethically possible. To do so, will be difficult but rewarding, and contribute to the credibility and scientific value of the trial results.

7. REFERENCES

1. Pocock SJ. Clinical trials. A practical approach. New York: Wiley, 1988.
2. Department of Health and Human Services, Food and Drug Administration. International Conference on Harmonisation; Guidance on Statistical Principles for Clinical Trials Availability. Federal Register, 63 (179), 1998: 49583-49598.
3. Food and Drug Administration. Guideline for Industry. Structure and Content of Clinical Study reports. FDA, 1996: at internet webside WWW.DFA.GOV /CDER/REGGUIDE.HTM/GUIDANCE DOCUMENTS.
4. Pocock SJ. Clinical trials. A practical approach. New York: Wiley, 1988, page 147.
5. Pocock SJ. Clinical trials. A practical approach. New York: Wiley, 1988, page 153.
6. Pocock SJ. Clinical trials. A practical approach. New York: Wiley, 1988, page 149.
7. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64: 191-199.
8. Demets DL, Ware JH. Group sequential methods in clinical trials with a one-sided hypothesis. Biometrika, 1980, 67: 651-660.
9. Wald A. Sequential Analysis. New York: Wiley, 1947.
10. Armitage P. Sequential Medical Trials. Oxford: Blackwell, 1975.
11. Whitehead J. The design and analysis of sequential clinical trials. Chichester: Ellis Horwood publishers, 1983.
12. Whitehead J. Planning and Evaluating Sequential Trials (PEST, version 3). Reading: University of Reading, 1998 (www.reading.ac.uk/mps/pest/pest.html)

CHAPTER 8

CONTROLLING THE RISK OF FALSE POSITIVE CLINICAL TRIALS

1. INTRODUCTION

Statistical hypothesis testing is much like gambling. If, with gambling once, your chance of a prize is 5%, then, with gambling 20 times, this chance will be close to 40%. The same is true with statistical testing of clinical trials. If, with one statistical test, your chance of a significant result is 5%, then after 20 tests, it will increase to 40%. This result is, however, not based on a true treatment effect, but, rather, on the play of chance. In current clinical trials, instead of a single efficacy-variable of one treatment, multiple efficacy-variables of more than one treatment are increasingly assessed. E.g., in 16 randomized controlled trials with positive results, published in the British Medical Journal (BMJ) in 2004 (Table 1), the numbers of primary efficacy-variables varied from 4 to 13. This phenomenon introduces the statistical problem of multiple comparisons and multiple testing, which increases the risk of false positive results, otherwise called type I errors. There is no consensus within the statistical community on how to cope with this problem. Also, the issue has not been studied thoroughly for every type of variable. Clinical trials rarely adjust their data for multiple comparisons. E.g., none of the above BMJ papers did. The current chapter briefly summarizes the main methods for control in order to further emphasize the importance of this issue, and it gives examples.

Table 1. Positive randomized controlled trials published in the BMJ in 2004

	Numbers of Primary Efficacy Variables	Smallest p-values	Positive Study after Bonferroni Adjustment
1. Schroter et al 328: 742-3	5	0.001	yes
2. Laurant et al 328: 927-30	12	0.006	no
3. Yudkin et al 328: 989-90	10	0.001	yes
4. Craig et al 328: 1067-70	6	0.030	no
5. Kalra et al 328: 1099-101	7	0.001	yes
6. Hilten et al 328: 1281-1	5	0.05	no
7. James et al 328: 1237-9	10	0.003	yes
8. Logan et al 328: 1372-4	6	0.01	no
9. Cairns S Smith et al 328: 1459-63	13	0.002	yes
10.Powell et al 329: 89-91	10	0.001	yes

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 107–111.

11.Henderson et al 329: 136-9	6	0.03	no
12.Collins et al 329: 193-6	4	0.03	no
13.Svendsen et al 329: 253-8	7	0.02	no
14.McKendry M 329: 258-61	9	0.001	yes
15.Van Staaij et al 329: 651-4	8	0.01	no
16.Norman et al 329: 1259-62	10	0.02	yes

2. BONFERRONI TEST

If more than two samples are compared in a clinical trial, multiple groups analysis of variance (ANOVA) is often applied for the analysis. E.g., three groups of patients were treated with different hemoglobin improving compounds with the following results:

	sample size	mean hemoglobin mmol / l	standard deviation mmol / l
Group1	16	8.725	0.8445
Group 2	10	10.6300	1.2841
Group 3	15	12.3000	0.9419

The F test produces a p-value < 0.01, indicating that a highly significant difference is observed between the three groups. This leads to the not-too-informative information that not all group means were equal. A question encountered is, which group did and which one did not differ from the others. This question involves the problem of multiple comparisons. As there are 3 different treatments, 3 different pairs of treatments can be compared: groups 1 versus 2, groups 1 versus 3, and groups 2 versus 3. The easiest approach is to calculate the Student's t-test for each comparison. It produces a highly significant difference at $p < 0.01$ between treatment 1 versus 3 with no significant differences between the other comparisons. This highly significant result is, however, unadjusted for multiple comparisons. If the chance of a falsely positive result is, e.g. α with one comparison, it should be 2α with two, and close to 3α with three comparisons. Bonferroni recommends to reject the null - hypothesis at a lower level of significance according to the formula

rejection p-value = α x 2 / k (k-1)
k = number of comparisons, α = agreed chance of falsely positive result (mostly 0.05)

In case of three comparisons the rejection p-value will be $0.05 \text{ x } \frac{2}{3(3-1)} = 0.0166$.

A p-value of 0.0166 is still larger than 0.01, and, so, the difference observed remains significant, but using a cut-off p-value of 0.0166, instead of 0.05, the difference is not highly significant anymore.

3. LEAST SIGNIFICANT DIFFERENCE TEST (LSD) TEST

As an alternative to the Bonferroni test a refined t-test, the least significant difference (LSD) test, can be applied. This refined t-statistic has n-k degrees of freedom, where n is the number of observations in the entire sample and k is the number of treatment groups. In the denominator of this refined t-test the usual pooled standard error (SE) is replaced with the pooled-within-group variance from the above mentioned F-test. For the application of the LSD procedure, it is essential to perform it sequentially to a significant F-test of the ANOVA procedure. So, if one chooses to perform the LSD procedure, one first calculates the ANOVA procedure and stops if it is not significant, and calculates the LSD test only if the F-test is statistically significant. The LSD test is largely similar to the Bonferroni-test, and yields with the above example a p-value close to 0.05. Like with Bonferroni, the difference is still significant, but not highly significant anymore.

4. OTHER TESTS FOR ADJUSTING THE P-VALUES

None of the 16 BMJ trials discussed in the introduction were adjusted for multiple testing. When we performed a Bonferroni adjustment of them, only 8 trials continued to be positive, while the remainder turned into negative studies. This does not necessarily indicate that all of these studies were truly negative. Several of them had more than 5 efficacy-variables, and, in this situation, the Bonferroni test is somewhat conservative, meaning that power is lost, and the risk of falsely negative results is raised. This is particularly so, if variables are highly correlated. A somewhat less conservative variation of the Bonferroni correction was suggested by Hochberg: if there are k primary values multiply the highest p-value with 1, the second-largest p-value with 2, the third largest with 3….., and the smallest p-value with k.[1]

Calculated p-values	reject null-hypothesis at
(1) largest p-value	$\alpha_1 = 0.05 \times 1 = 0.05$
(2) second largest p-value	$\alpha_2 = 0.05 \times 2 = 0.10$
(3) third largest p-value	$\alpha_3 = 0.05 \times 3 = 0.15$
(k) kth largest p-value	$\alpha_k = 0.05 \times k = \ldots.$

The mathematical arguments of this procedure goes beyond this paper. What happens is, that the lowest and highest p-values will be less different from one another. There are other less conservative methods, like Tukey's honestly significant difference (HSD) test, Dunnett's test, Student-Newman-Keuls test, and the Hotelling Q-square test. Most of them have in common that they produce their own test-statistics. Tables of significance levels are available in statistical software packages including SAS and SPSS.

5. COMPOSITE ENDPOINT PROCEDURES

A different solution for the multiple testing problem is to construct a composite endpoint of all of the efficacy-variables, and, subsequently, to perform a statistical analysis on the composite only. For example, it is reasonable to believe that statin treatment has a beneficial effect on total cholesterol (Tc), high density cholesterol (HDL), low density cholesterol (LDL), and triglycerides (Tg). We can perform a composite analysis of the four variables according to

Composite variable = $(T_c + HDL + LDL + Tg) / 4$

$$T_c = \frac{(T_c - \text{mean}(T_c))}{SDT_c} \text{ etc}$$

A simple t-test produces
Placebo: mean result composite variable= -0.23 (SD 0.59)
Statin: " " " " = 0.15 (SD 0.56)
$p = 0.006$

This p-value is lower than that obtained by a Bonferroni or LSD procedure. This is probably so, because of the added power provided by the positive correlation between the repeated observations in one subject. If no strong correlation between the variables is to be expected, the composite endpoint procedure provides power similar to that of the Bonferroni or LSD procedure.

Largely similar to the composite endpoint procedure are the index methods. If the efficacy-variables are highly correlated, because they more or less measure the same patient characteristic, then they be best replaced with their add-up sum. In this way the number of primary variables is reduced, and an additional advantage is that the standardized add-up sum of the separate variables is more reliable than the separate variables. E.g., the Disease Activity Score (DAS) for the assessment of patients with rheumatoid arthritis, including the Ritchie joint pain score, the number of swollen joints, and the erythrocyte sedimentation rate, is an example of this approach.[2]

6. NO ADJUSTMENTS AT ALL, AND PRAGMATIC SOLUTIONS

A more philosophical approach to the problem of multiple comparisons is to informally integrate the data, look for trends without judging one or two low p-values among otherwise high p-values as proof of a significant difference in the data. However, both the medical community and the investigators may be unhappy with this solution, because they want the hard data to provide unequivocal answers to their questions, rather than uncertainties. An alternative and more pragmatic solution could be the standard use of lower levels of significance to reject the null-hypothesis. For the statistical analysis of interim analyses, that suffer from the same risk of increased type I errors due to multiple testing, Pocock's recommendation to routinely use $p<0.01$

instead of $p<0.05$ has been widely adopted.[3] A similar rule could, of course, be applied to any multiple testing situation. The advantage would be that it does not damage the data, because the data remain undamaged. Moreover, any adjustments may produce new type I errors, particularly, if they are post-hoc, and not previously described in the study protocol.

7. CONCLUSIONS

Approaches to reducing the problem of multiple testing include (1) the Bonferroni, test, (2) the LSD method, (3) other less conservative, more rarely used methods like Tukey's honestly significant (HSD) method, Dunnett's test, Student-Newman-Keuls test, Hochberg's adjustment, and the Hotelling Q-square test. Alternative approaches to the problems of multiple testing include (4) the construct of composite endpoints, (5) no adjustment at all, but a more philosophical approach to the interpretation of the p-values, and (6) the replacement of the traditional 5% rejection level with a 1% rejection level or less.

Evidence-based medicine is increasingly under pressure, because clinical trials do not adequately apply to their target populations.[4-6] Many causes are mentioned. As long as the issue of multiple testing is rarely assessed in the analysis of randomized controlled trials, it can not be excluded as one of the mechanisms responsible. We recommend that the increased risk of false positive results should be taken into account in any future randomized clinical trial which assesses more than one efficacy-variable and / or treatment modality. The current chapter provides 6 possible methods for assessment.

8. REFERENCES

1. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75: 800-2.
2. Fuchs HA. The use of the disease activity score in the analysis of clinical trials in rheumatoid arthritis. J Rheumatol 1993; 20: 1863-6.
3. Pocock SJ. Clinical trials. A practical approach. New York, Wiley, 1988.
4. Furberg C. To whom do the research findings apply? Heart 2002; 87: 570-4.
5. Julius S. The ALHATT study: if you believe in evidence-based medicine. Stick to it. Hypertens 2003; 21: 453-4.
6. Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. Int J Clin Pharmacol Ther 2003; 41: 51-6.

CHAPTER 9

MULTIPLE STATISTICAL INFERENCES

1. INTRODUCTION

Clinical trials often assess the efficacy of more than one new treatment and often use many efficacy variables. Also, after overall testing these efficacy variables, additional questions about subgroups differences or about what variables do or do not contribute to the efficacy results, remain. Assessment of such questions introduces the statistical problem of multiple comparison and multiple testing, which increases the risk of false positive statistical results, and thus increases the type-I error risk. In the previous chapter six commonly-used methods for controlling the risk of this problem have been addressed. This chapter gives a more mathematical approach of the problem, and gives examples in which different methods are compared with one another.

2. MULTIPLE COMPARISONS

When in a trial three of more treatments are compared to each other, the typical first statistical analysis is to test the null hypothesis (H_0) of no difference between treatments versus the alternative hypothesis (H_a) that at least one treatment deviates from the others. Suppose that in the trial k different treatments are compared, then the null hypothesis is formulated as $H_0 : \vartheta_1 = \vartheta_2 = ... = \vartheta_k$, where ϑ_i is the treatment-effect of treatment i. When the efficacy variable is quantitative (and normally distributed), then 3 is the mean value. When the efficacy variable is binary (e.g. healthy or ill), then 3 is the proportion of positive (say healthy) patients. When the efficacy variable is of ordinal character, or is a survival time, 3 can have different quantifications. For the remainder of this paragraph we assume that the efficacy is quantitative and normally distributed, because for this situation the multiple comparison procedure has been studied thoroughestly.
Consider the randomized clinical trial comparing 5 different treatments for ejaculation praecox[1]: one group of patients received a placebo treatment (group 1), and the four other groups received different serotonin reuptake inhibitors (SSRI). The primary variable for evaluating the efficacy was the logarithmically transformed intravaginal ejaculation latency time (IELT) measured after six weeks of treatment. The null hypothesis in this trial was that there was no difference between the five groups of patients with respect to the mean of the logarithmically transformed IELT: $H_0 : \vartheta_1 = \vartheta_2 = \vartheta_3 = \vartheta_4 = \vartheta_5$. The summarizing data of this trial are listed in Table 1.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 113–122.

Table 1. Randomized clinical trial comparing 5 different treatments for ejaculation praecox[1]: one group of patients received a placebo treatment (group 1), and the four other groups received different serotonin reuptake inhibitors (SSRI). The primary variable for evaluating the efficacy was the logarithmically transformed intravaginal ejaculation latency time (IELT) measured after six weeks of treatment

Treatment	**sample size n**	**Mean x**	**Standard deviation S**
Placebo	9	3.34	1.14
SSRI A	6	3.96	1.09
SSRI B	7	4.96	1.18
SSRI C	12	5.30	1.51
SSRI D	10	4.70	0.78

The first statistical analysis was done by calculating the analysis of variance (ANOVA) table. The F-test for the testing the null hypothesis had value 4.13 with 4 and 39 degrees of freedom and p-value 0.0070. The within group sums of squares was 55.16 with 39 degrees of freedom, thus the mean squared error was S=1.41. Since the p-value was far below the nominal level of α =0.05, the null hypothesis could be rejected. This led to the not-too-informative conclusion that not all population averages were equal. A question immediately encountered is which one of the different population did and which one did not differ from each other. This question concerns the problem of multiple comparisons or post-hoc comparison of treatment groups.

The only way of finding out which one of the populations means differ from each other is to compare every treatment group with all of the other groups or with a specified subset receiving other treatments. When there are 5 different treatments, 5x4/2=10 different pairs of treatments can be compared. In general, when there are k treatments, k (k-1) / 2 different comparisons can be made.

The easiest approach to this question is to calculate the Student's t-test for each comparison of the groups i and j. This procedure may be refined by using in the denominator of the t-test the pooled-within-group variance S_w^2, as already calculated in the above F-test according to:

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{S_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}. \qquad (1)$$

This t-statistic has n-k degrees of freedom, where n is the total number of observations in the entire sample and k is the number of treatment groups. This procedure is called the "least significant difference" procedure (LSD procedure). For the application of the LSD procedure, it is essential to perform it sequentially to a significant F-test of the ANOVA procedure. So if one chooses to perform the LSD procedure, one first calculates the ANOVA procedure and stops if the F-test is non-significant, and calculates the LSD tests only when the F-test is statistically significant. The LSD procedure is largely similar to the Bonferroni-t-test for paired comparisons as explained in Chapter 2 section 3.

When the different treatment groups are compared without performing ANOVA first, or when you do so without the F-test being significant, then the problem of multiple comparisons is encountered. This means that when you make enough comparisons, the chance of finding a significant difference will be substantially larger than the nominal level of $\alpha = 0.05$: thus the risk of a type-I error will be (far) too large. There may be situations where we want to further the analysis all the same.

There are several ways, then, of dealing with the problem of an increased risk of type-I-error. The easiest method is to use the Bonferroni-correction, sometimes known as the modified LSD procedure. The general principle is that the significance level for the experiment, α_E is less than or equal to the significance level for each comparison, α_C, times the number of comparisons that are made (remember α is the chance of a type-I-error or the chance of finding a difference where there is none):

$$\alpha_E \leq \frac{k(k-1)}{2}\alpha_C \tag{2}$$

If $\alpha_E \leq 0.05$, then this level of α is maintained if , α_C is taken to be , divided by the number of comparisons:

$$\alpha_C = \alpha \frac{2}{k(k-1)} \tag{3}$$

When k is not too large, this method performs well. However, if k is large (k>5), then the Bonferroni correction is overconservative, meaning that the nominal significance level soon will be much lower than $\alpha = 0.05$ and loss of power occurs accordingly.

There are several alternative methods[2], but here we will discuss briefly three of them: Tukey's honestly significant difference (HSD) method, the Student-Newman-Keuls method, and the method of Dunnett. Tukey's HSD method calculates the test-statistic from the above equation (1), but determines the significance level slightly differently, by considering the distribution of the largest standardized difference $|x_i - x_j| / se\ (x_i - x_j)$. This distribution is somewhat more complex than that of the t-distribution or of the LSD procedure. A table of significance levels is available in all major statistical books as well as statistical software packages such as SAS and SPSS.[3,4] The HSD procedure

controls the maximum experiment wise error rate, and performs well in simulation studies, especially when sample sizes are unequal.

The Student-Newman-Keuls (SNK) procedure is a so-called multiple-stage or multiple range test. The procedure first tests the homogeneity of all k means at the nominal level α_k. When the homogeneity is rejected, then each subset of (k-1) means is tested for homogeneity at the nominal level α_{k-1}, and so on. It does so by calculating the studentized statistic in the above equation (1) for all pairs. The distribution of this statistic is again rather complex, and it depends on the degrees of freedom n-k (from ANOVA), on the number of comparisons that are made, and on α_k. The table of significance levels is likewise available in most statistical packages.

The conclusions of the SNK procedure critically depend on the order of the pair wise comparisons being made. The proper procedure is to compare first the largest mean with the smallest, then the largest with the second-smallest, and so on. An important rule is that if no significant difference exists between two means, it should be concluded that no difference exists between any means enclosed by the two, without further need of testing.

There are many multiple range tests[2], mainly differing in their use of the significance level α_k, and α_{k-1}. The Student-Newman-Keuls procedure uses $\alpha_k = \alpha = 0.05$, and therefore does not control the maximum experimentwise error rate.

Finally, there is a special multiple comparison procedure for comparing all active treatments to a control or placebo group. This is the Dunnett's procedure. For all treatments the studentized statistic of above equation (1) compared to the placebo group is calculated. In case of Dunnett's procedure, this statistic again has a complex distribution (many-one t-statistic) which depends on the number of active treatment groups, the degrees of freedom and a correlation term which depends on the sample sizes in each treatment group. Tables are likewise available in statistical packages. If sample sizes are not equal, it is important to use the harmonic mean of the sample sizes when calculating the significance of the Dunnett's test.

Most of the statistical packages compute common multiple range tests, and provide associated confidence intervals for the difference in means. In our trial comparing 4 SSRIs and placebo in patients with ejaculation praecox, we were interested in all of the possible comparisons between the five treatment groups. Since the ANOVA F-test was statistically significant, we applied the LSD procedure to find out which treatment differed significantly from each other. We found the following results. HSD procedure, the Bonferroni correction, and Dunnett's procedure of the same data were applied for control (Table 2).

Table 2. In the trial from Table 1 the investigators were interested in all of the possible comparisons between the five treatment groups. Since the ANOVA F-test was statistically significant, we applied the LSD procedure to find out which treatment differed significantly from each other. We found the following results. HSD procedure, the Bonferroni correction, and Dunnett's procedure of the same data were applied for control

		Difference	**P value**			
		Mean (SE)	**LSD**	**HSD**	**Bonferroni**	**Dunnett**
Placebo vs	A	-0.62 (0.63)	0.33	0.86	0.99	0.73
	B	-1.62 (0.60)	0.01	0.07	0.10	0.035
	C	-1.96 (0.52)	0.001	0.005	0.006	0.002
	D	-1.36 (0.55)	0.017	0.12	0.17	0.058
A vs	B	-1.00 (0.66)	0.14	0.56	0.99	
	C	-1.34 (0.60)	0.03	0.18	0.30	
	D	-0.74 (0.61)	0.24	0.75	0.99	
B vs	C	-0.34 (0.57)	0.56	0.98	0.99	
	D	0.26 (0.59)	0.66	0.99	0.99	
C vs	D	0.60 (0.51)	0.25	0.76	0.99	

SE=standard error.

The mean difference indicates the differences of the means of the groups as shown in Table 1. The standard error as calculated from the studentized statistic in the equation (1), and is required in order to construct confidence intervals. The critical values for the construction of such confidence intervals are supplied by appropriate tables for the HSD, and Dunnett's procedure, but are also calculated by most statistical software programs. In our case it is obvious that the LSD procedure provides the smallest p-values, and significant differences between SSRIs B, C and D and placebo results, as well as between A and C results. When using the Bonferroni test or the HSD procedure, only SSRI C is significantly different from placebo. Dunnett's test agrees with the LSD procedure with respect to the differences of the SSRIs compared to placebo, but has no information on the differences between the SSRIs.

There is no general consensus on what post-hoc test to use or when to use it; as the statistical community has not yet reached agreement on this issue. The US Food and Drug Agency suggests in its clinical trial handbook for in house usage to describe in the study protocol the arguments for using a specific method, but refrains from

making any preference. We have a light preference for calculating an overall test first such as is done with ANOVA, and subsequently proceed with the LSD test.
Unfortunately, so far multiple comparisons methods have not been developed much for discrete, ordinal and censored data. When dealing with such data, it is best to perform first an overall test by chi-square, Kruskal-wallis or logrank methods, and afterwards perform pairwise comparisons with a Bonferroni correction.
Whatever method for multiple comparisons, its use or the lack of its use should be discussed in the statistical analysis, and preferably be specified in the analysis plan of the study protocol.

3. MULTIPLE VARIABLES

Most clinical trials use several, and sometimes many, endpoints to evaluate the treatment efficacy. The use of significance tests separately for each endpoint comparison increases the risk of a type-I error of finding a difference where there is none. The statistical analysis should reflect awareness of this very problem, and in the study protocol the use or non-use of statistical adjustments or their lack must be explained. There are several ways of handling this problem of multiple testing.

I The most obvious way is to simply reduce the number of endpoint parameters otherwise called primary outcome variable. Preferably, we should include one primary parameter, usually being the variable that provides the most relevant and convincing evidence of the primary objective of the trial. The trial success is formulated in terms of results demonstrated by this very variable, and prior sample size determination is also based on this variable. Other endpoint variables are placed on a lower level of importance and are defined secondary variables. The secondary variable results may be used to support the evidence provided by the primary variable.

It may sometimes be desirable to use two or more primary variables, each of which sufficiently important for display in the primary analysis. The statistical analysis of such an approach should be carefully spelled in the protocol. In particular, it should be stated in advance what result of any of these variables is least required for the purpose of meeting the trial objectives. Of course, if the purpose of the trial is to demonstrate a significant effect in two or more variables, then there is no need for adjustment of the type-I error risk, but the consequence is that the trial fails in its objectives if one of these variables do not produce a significant result. Obviously, such a rule enhances the chance of erroneously negative trials, in a way similar to the risk of negative trials due to small sample sizes.

II A different more philosophical approach to the problem of multiple outcome variables is to look for trends without judging one or two low P-values among otherwise high P-values as proof. This requires discipline and is particularly efficient when multiple measurements are performed for the purpose of answering one single question, e.g., the benefit to health of a new drug estimated in terms of effect on mortality in addition to a number of morbidity variables. There is nothing wrong with

this practice. We should not make any formal correction for multiple comparisons of this kind (see also Chapter 1, section 1). Instead, we should informally integrate all the data before reaching a conclusion.

III An alternative way of dealing with the multiple comparison problem when there are many primary variables, is to apply a Bonferroni correction. This means that *the p-value of every variable is multiplied by the number of endpoints k.* This ensures that if treatments were truly equivalent, the trial as a whole will have less than a 5% chance of getting any p-value less than 0.05; thus the overall type-I error rate will be less than 5%.

IV The Bonferroni correction, however, is not entirely correct when multiple comparisons are dependent of each other (multiple comparisons in one subject cannot be considered independent of each other, compare chapter 2, section 3, for additional discussion of this issue). Also the Bonferroni correction is an overcorrection in case of larger numbers of endpoints, particularly when different endpoints are (highly) correlated. A somewhat more adequate variation of the Bonferroni correction, was suggested by Hochberg.[5] *When there are k primary values, the idea is to multiply the largest p-value with 1, the second-largest p-value with 2, the third largest p-value with 3, ..., and the smallest p-value with k.* We do not attempt to explain the mathematical arguments of this procedure, but conclude that lowest and highest –values will be less different from each other. In practice, Hochberg's procedure is frequently hardly less conservative than is the Bonferroni correction.

V An further alternative for analyzing two or more primary variables is to design a summary measure or composite variable. With such an approach endpoint and primary variables must, of course, be assessed in advance, and the algorithm to calculate the composite must also be specified a priori. Since in this case primary variables are reduced to one composite, there is no need to make adjustments to salvage the type-I error rate. For the purpose of appropriate composite variables there are a few sensible rules to bear in mind:

- Highly correlated variables, measuring more or less the same patient characteristic can best be replaced by their average. In this way the number of primary variables is reduced, and an additional advantage is that the mean is more reliable than both single measurements.
- When the variables have different scales (e.g. blood pressure is measured in mm Hg units, and cholesterol in mmol/L units), the composite variables are best calculated as standardized variables. This means that the overall mean is subtracted from each measurement and that the resulting difference is divided by the overall standard deviation. In this way all variables will have zero mean and unit standard deviation in the total sample.

Well-known examples of composite variables are rating scales routinely used for the assessment of health-related quality of life, as well as disease-activity-scales (e.g., the disease activity scale of Fuchs for patients with rheumatoid arthritis, DAS[6]). The DAS is a composite based on the Ritchie joint pain score, the number of swollen joints, and, in addition, the erythrocyte sedimentation rate:

$$DAS = 0.53938\sqrt{ritchie\ index} + 0.06465(number\ of\ swollen\ joints) +$$

$$0.330\ln(erythocyte\ sedimentation\ rate) + 0.224.$$

For the statistical analysis of a composite variable, standard methods may be used without adjustments. Lauter[7] showed that the statistical test for the composite has 5% type-I error rate. He also showed that such a statistical test is especially sensitive when each endpoint variable has more or less the same individual p-value, but that it has little sensitivity when one endpoint variable is much more significant than others.
We applied these methods to a clinical trial of patients with atherosclerosis comparing two-year placebo versus pravastatin medication.[8] The efficacy of this medication was evaluated by assessing the change of total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides. The mean changes and standard deviations (mmol/L) are given in Table 3, while also the uncorrected p-values, and the corrected p-values according to Bonferroni and Hochberg are reported.

Table 3. Clinical trial of patients with atherosclerosis comparing two-year placebo versus pravastatin medication.[8] The efficacy of this medication was evaluated by assessing the change of total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides. The mean changes and standard deviations (mmol/L) are given, while also the uncorrected p-values, and the corrected p-values according to Bonferroni and Hochberg are reported

Change of:	**Placebo (n=31)**	**Pravastatin (n=48)**	**P***	**P#**	**P@**
Total cholesterol decrease	-0.07 (0.72)	0.25 (0.73)	0.06	0.24	0.11
HDL cholesterol increase	-0.02 (0.18)	0.04 (0.12)	0.07	0.28	0.11
LDL cholesterol decrease	0.34 (0.60)	0.59 (0.65)	0.09	0.36	0.11
Triglycerides increase	0.03 (0.65)	0.28 (0.68)	0.11	0.44	0.11

* p-value of Student's t-test; # Bonferroni corrected p-value; @ p-value corrected using Hochberg's methods.

It is obvious that none of the changes are statistically significant using a standard t-test, but it is also clear that all four efficacy variables have a treatment difference that points in the same direction, namely of a positive pravastatin effect. When correcting for multiple testing, the p-values are nowhere near statistical significance. A composite variable of the form z = (total cholesterol + HDL + LDL + triglycerides)/4, where the four lipid measurements are standardized, however, did show statistically significant results: the mean of Z in the placebo group was -0.23 (SD 0.59), and the mean of Z in the pravastatin group was 0.15 (SD 0.56), different $p<0.01$, and so, it is appropriate to conclude that pravastatin significantly reduced the composite variable.

VI Finally, there are several multivariate methods to perform an overall statistical test for which the type-I error risk equals 5%. Equivalently to the situation comparing many different treatment groups, one might argue that the overall test controls the type-I error, and that subsequently to the overall test, one can perform t-tests and the like without adjustment to explore which variables show significant differences. For comparing two treatment groups on several (normally distributed) variables, one may use Hotelling's T-square, which is the multivariate generalization of the Student's t-test. Other methods to compare different groups of patients on several variables are discriminant analysis, variants of principal components analysis and multinominal logistic regression. The discussion of these methods falls outside the scope of this chapter. It suffices to remark that Hotelling's T-square and the other multivariate methods are readily available through most statistical packages.

4. CONCLUSIONS

Multiple group comparison and multiple variable testing is a very common problem when analyzing clinical trials. There is no consensus within the statistical community on how to cope with these problems. It is therefore essential that awareness of the existence of these problems is reflected in the study protocol and the statistical analysis.

5. REFERENCES

1. Waldinger MD, Hengeveld MW, Zwinderman AH, Olivier B (1998). Effect of SSRI antidepressants on ejaculation: A double-blind, randomized, placebo-controlled study with fluoxetine, fluvoxamine, paroxetine, and sertraline. Journal of Clinical Psychopharmacology, 18 (4): 274-281.
2. Multiple comparisons boek, Edition University of Leiden , Neth, 1999.
3. SAS Statistical Software, 1998.
4. SPSS Statistical Software, Chicago, IL, 1996.
5. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988, 75: 800-802.
6. Fuchs HA. The use of the disease activity score in the analysis of clinical trials in rheumatoid arthritis. J Rheumatol, 1993, 20(11): 1863-6.

7. Lauter J. Exact t and F-tests for analyzing studies with multiple endpoints. Biometrics 1996, 52: 964-970.
8. Jukema JW, Bruschke AV, Van Boven AJ, Zwinderman AH, et al. Effects of lipid lowering by pravastatin on the regression of coronary artery disease in symptomatic men. Circulation 1995; 91: 2528-40.

CHAPTER 10

THE INTERPRETATION OF THE P-VALUES

1. INTRODUCTION

In randomized controlled trials, prior to statistical analysis, the data are checked for outliers and erroneous data. Data-cleaning is defined as deleting-the-errors / maintaining-the-outliers. Statistical tests are, traditionally, not very good at distinguishing between errors and outliers. However, they should be able to point out main endpoint results that are closer to expectation than compatible with random sampling. E.g., a difference from control of 0.000 is hardly compatible with random sampling. As it comes to well-balanced random sampling of representative experimental data, nature will be helpful to provide researchers with results close to perfection.
However, because biological processes are full of variations, nature will never allow for 100 per cent perfection. Statistical distributions can account for this lack of perfection in experimental data sampling, and provide exact probability levels of finding results close to expectation.

2. RENEWED ATTENTION TO THE INTERPRETATION OF THE PROBABILITY LEVELS, OTHERWISE CALLED THE P-VALUES

The p-values tell us the chance of making a type I error of finding a difference where there is none. Generally, a cut-off p-value of 0.05 is used to reject the null-hypothesis (H_0) of no difference. In the seventies exact p-values were laborious to calculate, and they were, generally, approximated from statistical tables, in the form of $p<0.01$ or $0.05<p<0.10$ etc. In the past decades with the advent of computers the job became easy.[1-4] Exact p-values such as 0.84 or 0.007 can now be calculated fast and accurately. This development lead to a renewed attention to the interpretation of p-values. In business statistics[5,6], the 5% cut-off p-value has been largely abandoned and replaced with exact p-values used for making decisions on the risk business men are willing to take, mostly in terms of costs involved. In medicine, the cut-off p-values have not been completely abandoned, but broader attention is given to the interpretation of the exact p-values, and rightly so, because they can tell us a number of relevant things in addition to the chance of making type I errors. In the current chapter standard and renewed interpretations of p-values are reviewed as far as relevant to the interpretation of clinical trials and evidence-based medicine.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 123–135.

3. STANDARD INTERPRETATION OF P-VALUES

Statistics gives no certainties, only chances. What chances? Chances that hypotheses are true/untrue (we accept 95% truths). What hypotheses? E.g., no difference from a 0 effect, a real difference from a 0 effect, worse than a 0 effect. Statistics is about estimating such chances / testing such hypotheses. Trials often calculate differences between test treatment and control (for example, standard treatment, placebo, baseline), and, subsequently, test whether the difference-between-the-two is different from 0.

Important hypotheses are Hypothesis 0 (H_0, i.e., no difference from a 0 effect), and Hypothesis 1 (H_1, the alternative hypothesis, i.e., a real difference from a 0 effect). What do these two hypotheses look like in graph? Figure 1 gives an example.

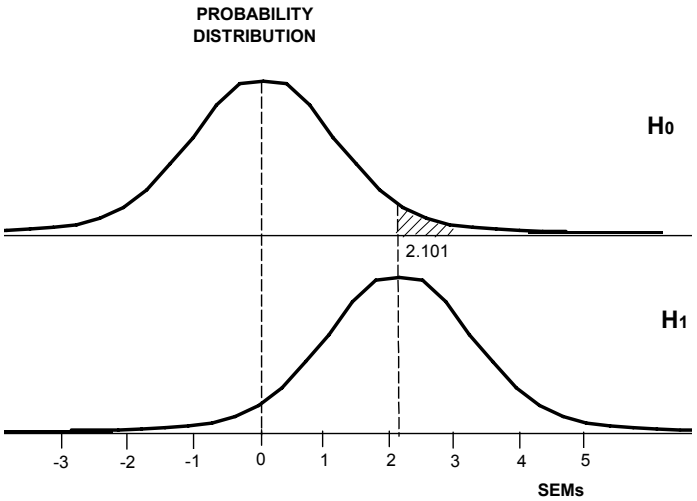


Figure 1. Null-hypothesis and alternative hypothesis of a parallel group study of two groups n=10 (18 degrees of freedom).

-H_1= graph based on the data of our trial (mean ± standard error (SEM) = 2.1 ± 1).
-H_0 = same graph with mean 0 (mean ± SEM= 0 ± 1).
-Now we make a giant leap from our data to the population from which the sample was taken (we can do so, because our data are supposed to be representative of the population).

-H_1 = also summary of means of many trials similar to ours (if we repeated trial, difference would be small, and distribution of means of many such trials would look like H_1.
-H_0 = summary of means of many trials similar to ours, but with overall effect 0 (our mean not 0 but 2.1. Still, it could be an outlier of many studies with an overall effect of 0.

-So, we should think of H_0 and H_1 as summaries of means of many trials.
-If hypothesis 0 is true, then mean of our study is part of H_0.
-If hypothesis 1 is true, then mean of our study is part of H_1.
-We can't prove anything, but we can calculate the chance of either of these possibilities.
-A mean result of 2.1 is far distant from 0:

Suppose it belongs to H_0.
Only 5% of the H_0 trials >2.1 SEM distant from 0.
The chance that it belongs to H0 is < 5%.
We reject this possibility if probability is <5%.

Suppose it belongs to H_1.
50% of the H_1 trials >2.1 SEM distant from 0. These 50% cannot reject null hypothesis, only the remainder, here also 50%, can do so.

Conclude here if H_0 is true, we have < 5% chance to find it, if H_1 is true, we have 50% chance to find it.
Or in statistical terms: we reject null hypothesis of no effect at $p < 0.05$ and with a statistical power of 50%.

Obviously, a p-value of < 0.05 does not indicate a true effect, and allows for very limited conclusions[7,8]:

(1) < 5% chance to find this result if H_0 is true (H_0 is probably untrue, and so, this statement does not mean too much anyway);
(2) only 50% chance to find this result if H_1 is true.

The conclusions illustrate the uncertainties involved in H_0 - testing. With lower p-values, better certainty is provided, e.g., with $p < 0.01$ we have around 80% chance to find this result if H_1 were true, with $p < 0.001$ even 90 %. However, even then, the chance of a type II error of finding no difference where there is one is still 10 %.
Also, we must realize that the above conclusions are appropriate only if

(3) the data follow a normal distribution, and
(4) they follow exactly the same distribution as that of the population from which the sample was taken.

4. COMMON MISUNDERSTANDINGS OF THE P-VALUES

The most common misunderstanding while interpreting the p-values is the concept that the p-value is actually the chance that the H_0 is true, and, consequently, that $p > 0.05$ means H_0 is true. Often, this result, expressed as "not significantly different from zero", is then reported as documented proof that the treatment had no effect. The distinction between demonstrating that a treatment had no effect and failing to demonstrate that it did have an effect, is subtle but very important, because the latter may be due to inadequate study methods or lack of power rather than lack of effect. Moreover, in order to assess whether the H_0 is true, null-hypothesis testing can never give the answer, because this is not the issue. The only issue here is: H_0 is rejected or not, no matter if it is true or untrue. To answer the question whether no-difference-in-the-data is true, we need to follow a different approach: similarity testing. With similarity (otherwise called equivalence)-testing the typical answer is: similarity is or is not demonstrated, which can be taken synonymous for no-difference-in-the-data being true or not (see also chapter 4).

5. RENEWED INTERPRETATIONS OF P-VALUES, LITTLE DIFFERENCE BETWEEN P = 0.06 AND P = 0.04

H_0 is currently less dogmatically rejected, because we believe that such practice mistakenly attempts to express certainty of statistical evidence in the data. If the H_0 is rejected, it is also no longer concluded that there is no difference in the data. Instead, we increasingly believe that there is actually little difference between $p = 0.06$ and $p = 0.04$. Like with business statistics clinicians now have the option to use p-values for an additional purpose, i.e., for making decisions about the risks they are willing to take.

Also an advantage of the exact p-value approach is the possibility of more refined conclusions from the research: instead of concluding significantly yes / no, we are able to consider levels of probabilities from very likely to be true, to very likely to be untrue.[9] The p-value which ranges from 0.0 to 1.0 summarizes the evidence in the data about H_0. A large p-value such as 0.55 or 0.78 indicates that the observed data would not be unusual if H_0 were true. A small p-value such as 0.001 denotes that these data would be very doubtful if H_0 were true. This provides strong support against H_0. In such instances results are said to be significant at the 0.001 level, indicating that getting a result of this size might occur only 1 out of 1000 times.

Exact p-values are also increasingly used for comparing different levels of significance. The drawback of this approach is that sampled frequency distributions are approximations, and that it can be mathematically shown that exactly calculated p-values are rather inaccurate.[10] However, this drawback is outweighed by the advantages of knowing the p-values especially when it gets to extremes.[11]

6. THE REAL MEANING OF VERY LARGE P-VALUES LIKE P > 0.95

Let us assume that in a Mendelian experiment the expected ratio of yellow-peas / green-peas = 1 / 1. A highly representative random sample of n=100 might consist of 50 yellow and 50 green peas. However, the larger the sample the smaller the chance to find exactly fifty/fifty. The chance of exactly 5000 yellow / 5000 green peas or even the chance of a result very close to this result is, due to large variability in biological processes, almost certainly zero.

Statistical distributions like the chi-square distribution can account for this lack of perfection in experimental data sampling, and provide exact probability levels of finding results close to "expected". Chi-squares curves are skewed curves with a lengthy right-end (Figure 2).

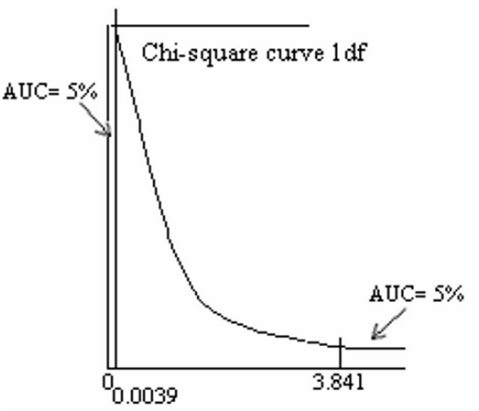


Figure 2. Probability of finding χ^2 value >3.841 is < 0.05, so is probability of finding a χ^2 value < 0.0039.

AUC = area under the curve; df = degree of freedom.

We reject the null-hypothesis of no difference between "expected and observed", if the area under curve (AUC) on the right side of the calculated chi-square value is < 5% of the total AUC. Chi-square curves do, however, also have a short left-end which ends with a chi-square value of zero. If the chi-square value calculated from our data is close to zero, the left AUC will get smaller and smaller, and as it becomes <5% of the total AUC, we are equally justified not to accept the null hypothesis as we are with large chi-square values. E.g., in a sample of 10,000 peas, you might find 4997 yellow and 5003 green peas. Are these data representative for a population of 1/1 yellow/green peas? In this example a chi-square value of < $3.9\ 10^{-3}$ indicates that the left AUC is <5% and, so, we have a probability <5% to find it (Table 1).[12]

Table 1. χ2 table: 7 columns of χ^2 values, upper two rows areas under the curve (AUCs) of left and right end of χ^2 curves, left column: adjustments for degrees of freedom (dfs)

AUC left end	.0005	.001	.005	.01	.025	.05	.10
AUC right end	.9995	.999	.995	.99	.975	.95	.90
degrees of freedom							
1	.0000004	.0000016	.000039	.00016	.00091	.0039	.016
2	.00099	.0020	.010	.020	.051	.10	.21
3	.015	.024	.072	.12	.22	.35	.58
4	.065	.091	.21	.30	.48	.71	1.06
5	.6	.21	.41	.55	.83	1.154	1.61

Chi-square value is calculated according to:
(Observed yellow-Expected yellow)^2 = (4997-5000)^2 : 5000 to standardize = 1.8 10^{-3}
(Observed green -Expected green)^2 = (5003-5000)^2 : 5000 to standardize = 1.8 10^{-3}
chi-square (1 degree of freedom) = 3.6 10^{-3}
This result is smaller than 3.9 10^{-3} and, thus, it is so close to what was expected that we can only conclude that we have < 5% probability to find it. We have to scrutinize these results, and must consider and examine the possibility of inadequate data improvement. The above example is actually based on some true historic facts (Mendel indeed improved his data).[13]

7. P-VALUES LARGER THAN 0.95, EXAMPLES (TABLE 2)

We searched for main endpoint p-values close to 0.95 in randomized controlled trials published in recent issues of the Lancet and the New England Journal of Medicine, and found four studies. Table 2 gives a summary. All of these studies aimed at demonstrating similarities rather than differences. Indeed, as can be observed, proportions of patients with events in the treatment and control groups were very similar. E.g., the percentages in treatment and control groups of patients with sepsis were 1.3 % and 1.3 % (study 1, Table 2), and of patients with cardiovascular events 79.2 % and 79.8 % (study 5, Table 2). The investigators of the studies calculated p-values from $p > 0.94$ to $p > 0.995$, which, according to the chi-square table (Table 1), would provide left-end p-values between ≤ 0.06 and ≤ 0.005. This would mean, that, for whatever reason, these data were probably not completely random. Unwarranted exclusion of, otherwise, appropriate outliers is one of the possible explanations.

Table 2. Study data with p-values close to 0.95, as published in recent Lancet and N Engl J Med issues

Result(numbers)	results(%)	Sample size requirement	alpha-level	P-values
Ref 14 107/6264 vs 107/6262	1.7 vs 1.7	yes	0.05	> 0.995
Ref 15 88/965 vs 84/941	9.1 vs 8.9	yes	0.05	> 0.95
Ref 15 13/965 vs 12/941	1.3 vs 1.3	yes	0.05	> 0.95
Ref 16 214/1338 vs 319/2319	15.9 vs 13.8	yes	0.05	> 0.99
Ref 17 285/360 vs 1087/1363	79.2 vs 79.8	yes	0.05	> 0.94

1. Proportions of patients with heart infarction in patients with diastolic blood pressure 80 < ... < 85 vs < 80 mm Hg. 2. Proportion of patients with arrhythmias in patients with standard perioperative treatment vs Swann-Ganz catheter-guided perioperative treatment. 3. Proportion of patients with sepsis in patients with standard perioperative treatment vs Swann-Ganz catheter-guided perioperative treatment. 4. Proportions of patients with cardiovascular events in patients with LDL-cholesterol, < 3.5 mmol/l vs 3.5 < ... < 4.5 mmol/l. 5. Proportions of patients with cardiovascular events in patients with LDL-cholesterol < 2.6 mmol/l vs > 3.4 mmol/l. Alpha = type I error, vs= versus.

8. THE REAL MEANING OF VERY SMALL P-VALUES LIKE P < 0.0001

Statistics gives no certainties, only chances. A generally accepted concept is “the smaller the p-value the better reliable the results”. This is not entirely true with current randomized controlled trials. First, randomized controlled trials are designed to test small differences. A randomized controlled trial with major differences between old and new treatment is unethical because half of the patients have been given an inferior treatment.

Second, they are designed to confirm prior evidence. For that purpose, their sample size is carefully calculated. Not only too small but also too large a sample size is considered unethical and unscientific, because negative studies have to be repeated and a potentially inferior treatment should not be given to too many patients. Often in the study protocol a statistical power of 80% is agreed, corresponding with a p-value of approximately 0.01.

The ultimate p-value may then be a bit larger or smaller. However a p-value of > 0.05 will be rarely observed, because current clinical trials are confirmational and, therefore, rarely negative. Also a p-value much smaller than 0.01 will be rarely observed, because it would indicate that either the power assessment was inadequate (the study is overpowered) or the data have been artificially improved. With p = 0.0001 we have a peculiar situation. In this situation the actual data can not only reject the null-hypothesis, but also the hypothesis of significantly better Thus, a p-value < 0.0001, if the power was set at 80%, does not completely confirm its prior expectations and must be scrutinized for data improvement. (This issue is explained more in detail in the next chapter).

9. P-VALUES SMALLER THAN 0.0001, EXAMPLES (TABLE 3)

Table 3 gives an overview of five published studies with main endpoint p-values < 0.0001. All of these studies were published in the first 6 issues of the 1992 volume of the New England Journal of Medicine. It is remarkable that so many overpowered studies were published within 6 subsequent months of a single volume, while the same journal published not any study with p-values below 0.001 in the past 4 years' full volumes. We do not know why, but this may be due to the journal's policy not to accept studies with very low p-values anymore. In contrast, many other journals including the Lancet, Circulation, BMJ, abound with extremely low p-values. It is obvious that these journals still believe in the concept "the lower the p-value, the better reliable the research". The concept may still be true for observational studies. However, in confirmational randomized controlled trials, p-values as low as 0.0001 do not adequately confirm prior hypotheses anymore, and have to be checked for adequacy of data management.

Table 3. Study data with p-values as low as < 0.0001, published in the first 6 issues of the 1992 volume of the N Engl J Med. In the past 4 years p-values smaller than p<0.001 were never published in this journal

	Result	Sample size requirement	alpha-level	P-values
Ref 18	+0.5 vs +2.1%	yes	0.05	< 0.0001
Ref 18	–2.8 vs +1.8 %	yes	0.05	< 0.0001
Ref 19	11 vs 19 %	no	0.05	< 0.0001
Ref 20	r = - 0.53	no	0.05	< 0.0001
Ref 21	213 vs 69	no	0.05	< 0.0001

1. Duration exercise in patients after medical therapy vs percutaneous coronary angioplasty. 2. Maximal double product (systolic blood pressure times heart rate) during exercise in patients after medical treatment vs percutaneous coronary angioplasty. 3. Erythromycin resistance throat swabs vs pus samples. 4. Correlation between reduction of epidermal pigmentation during treatment and baseline amount of pigmentation. 5. Adverse reactions of high vs non-high osmolality agents during cardiac catheterization. Alpha = type I error, vs = versus.

10. DISCUSSION

In 1948 the first randomized controlled trial was published by the BMJ.[22] Until then, observations had been mainly uncontrolled. Initially, trials were frequently negative due to little sensitivity as a consequence of too small samples, and inappropriate hypotheses based on biased prior data. Nowadays, clinical trials are rarely negative, and they are mainly confirmational rather than explorative. This has consequences for the p-values that can be expected from such trials. Very low p-values like $p<0.0001$ will be rarely encountered in such trials, because it would mean that the study was overpowered and should have had a smaller sample size. Also very large p-values like $p > 0.95$ will be rare, because they would indicate similarities closer than compatible with a normal distribution of random data samples.

We should emphasize that the above-mentioned interpretation of very low / high p-values is only true within the context of randomized controlled trials. E.g., unrandomized observational data can easily produce very low and very high p-values, and there is nothing wrong with that. Also the above interpretation is untrue in clinical trials that test multiple endpoints rather than a single main endpoint or a single composite endpoint. Clinical trials testing multiple rather than single endpoints, often do so for the purpose of answering a single question, e.g., the benefit of health of a new drug may be estimated by mortality in addition to various morbidity variables. If investigators test many times, they are apt to find differences, e.g., 5 % of the time, but this may not be due to significant effects but rather to chance. In this situation, one should informally integrate all of the data before reaching conclusions, and look for the trends in the data without judging one or two low p-values, among otherwise high p-values, as proof (see also the chapters 7 and 8).

In the present chapter, for the assessment of high p-values, the chi-square test is used, while for the assessment of low p-values the t-test is used. Both tests are, however, closely related to one another, and like other statistical tests, including the F-test, regression analysis, and other tests based on normal distributions. The conclusions drawn from our assessments are, therefore, equally true for alternative statistical tests and data.

We should add that the nominal p-values have to be interpreted with caution in case of multiple testing, as already discussed in the previous two chapters. A not yet mentioned but straightforward way to correct this is to calculate an E-value, i.e. the product of the p-value and the number of tests.

11. RECOMMENDATIONS

P-values < 0.0001 will be rarely encountered in randomized controlled clinical trials, because it would mean that the study is overpowered and should have had a smaller sample size. Also p-values > 0.95 will be rare, because they would indicate similarities closer than compatible with a normal distribution of random samples. It

would seem appropriate, therefore, to require investigators to explain such results, and to consider rejecting the research involved. So far, in randomized controlled trials the null-hypothesis is generally rejected at $p < 0.05$. Maybe, we should consider rejecting the entire study if the main endpoint p-values are > 0.95 or < 0.0001.

The concept of the p-value is notoriously poorly understood. Some physicians even comfortably think that the p-value is a measure of effect.[23] When asked whether a drug treatment worked, their typical answer would be: "Well, p is less than 0.05, so I guess it did". The more knowledgeable among us know that p stands for chance (probability = p), and that there must be risks of errors. The current paper reviews the standard as well as renewed interpretations of the p-values, and was written for physicians accepting statistical reasoning as a required condition for an adequate assessment of the benefits and limitations of evidence-based medicine.

Additional points must be considered when interpreting the p-values. In the first place, the interpretation of low p-values is different in studies that test multiple endpoints rather than a single main endpoint or a single composite endpoint. Studies testing multiple rather than single endpoints, often do so for the purpose of answering a single question, e.g., the benefit of health of a new drug may be estimated by mortality in addition to various morbidity variables. If investigators test many times, they are apt to find differences, e.g., 5% of the time, but this may not be due to significant effects but rather to chance. In this situation, one should informally integrate all of the data before reaching conclusions, and look for the trends in the data without judging one or two low p-values, among otherwise high p-values, as proof.

Special attention in this respect deserves the issue of multiple low-powered studies. One might consider this situation to be similar to the above one, and conclude that such studies be similarly integrated. Actually, this is one of the concepts of the method of meta-analysis. Second, the point of one sided testing versus two-sided testing must be considered. Studies testing both ends of a normal frequency distribution have twice the chance of finding a significant difference compared to those testing only one end. If our research assesses whether there is any difference in the data, no matter in what direction, either the positive or the negative one, then we have a two-sided design and the p-values must doubled. It is then, consequently, harder to obtain a low p-value.

Recommendations regarding the interpretation of main-endpoint-study p-values either two-sided or not, include the following.

1. $P < 0.05$ gives a conditional probability: H0 can be rejected on the limitations/assumptions that (1) we have up to 5% chance of a type I error of finding a difference where there is none, (2) we have 50% chance of a type II error of finding no difference where there is one, (3) the data are normally distributed, (4) they follow exactly the same distribution as that of the population from which the sample was taken.
2. A common misunderstanding is the concept that the p-value is actually the chance that H_0 is true, and, consequently that a $p > 0.05$ indicates a significant similarity in the data. $P > 0.05$ may, indeed, indicate similarity. However, also a study-sample too small or study design inadequate to detect the difference must be considered.
3. An advantage of the exact p-values is the possibility of more refined conclusions from the research: instead of concluding significantly yes/no, we are able to

consider levels of probabilities from very likely to be true, to very likely to be untrue.

4. $P > 0.95$ suggests that the observed data are closer to expectation than compatible with a Gaussian frequency distribution, and such results must, therefore, be scrutinized.
5. A $p < 0.0001$, if power was set at 80%, does not completely confirm the prior expectations of the power assessment. Therefore, such results must be scrutinized.

12. CONCLUSIONS

The p-values tell us the chance of making a type I error of finding a difference where there is none. In the seventies exact p-values were laborious to calculate, and they were, generally, approximated from statistical tables, in the form of $p < 0.01$ or $0.05 < p < 0.10$ etc. In the past decades with the advent of computers it became easy to calculate exact p-values such as 0.84 or 0.007. The cut-off p-values have not been completely abandoned, but broader attention is given to the interpretation of the exact p-values. The objective of this chapter was to review standard and renewed interpretations of p-values:

1. Standard interpretation of cut-off p-values like $p < 0.05$.

The null-hypothesis of no difference can be rejected on the limitations/assumptions that (1) we have up to 5% chance of a type I error of finding a difference where there is none, (2) we have 50% chance of a type II error of finding no difference where there is one, (3) the data are normally distributed, (4) they follow exactly the same distribution as that of the population from which the sample was taken.

2. A common misunderstanding of the p-value.

It is actually the chance that the null-hypothesis is true, and, consequently that a $p > 0.05$ indicates a significant similarity in the data. $P > 0.05$ may, indeed, indicate similarity. However, a study-sample too small or study design inadequate to detect the difference must be considered.

3. Renewed interpretations of the p-values.

Exact p-values enable to more refined conclusions from the research than cut-off levels: instead of concluding significantly yes/no, we are able to consider levels of probabilities from very likely to be true, to very likely to be untrue. Very large p-values are not compatible with a normal Gaussian frequency distribution, very small p-values do not completely confirm prior expectations. They must be scrutinized, and may have been inadequately improved.

13. REFERENCES

1. SAS. www.sas.com
2. SPSS. www.spss.com
3. S-plus. www.splus.com
4. Stata. www.stata.com
5. Levin RI, Rubin DS. P-value. In: Statistics for management. Eds Levin RI and Rubin DS, Prentice-Hall, New Jersey, 1998, pp 485-496.

6. Utts JM. P-value. In: Seeing through statistics. Ed Utts JM, Duxbury Press, Detroit, 1999, pp 375-386.
7. Cleophas TJ, Zwinderman AH, Cleophas AF. P-values. Review. Am J ther 2004; 11: 317-322.
8. Cleophas TJ, Zwinderman AH, Cleophas AF. P-values, beware of the extremes. Clin Chem Lab Med 2004; 42: 300-305.
9. Michelson S, Schofield T. p-values as conditional probabilities. In: The biostatistics cookbook. Eds Michelson S and Schofield T, Kluwer Academic Publishers, Boston, 1996, pp 46-58.
10. Petrie A, Sabin C. Explanation of p-values. In: Medical statistics at glance. Eds Petrie A and Sabin C, Blackwell Science Oxford UK, 2000, pp 42-45.
11. Matthews DE, Farewel VT. P-value. In: Using and understanding medical statistics. Eds Matthews DE and Farewell VT, Karger, New York, 1996, pp15-18.
12. Riffenburgh RH. P-values. In: Statistics in Medicine. Ed Riffenburgh RH, Academic Press, San Diego, 1999, pp 95-96, 105-106.
13. Cleophas TJ, Cleophas GM. Sponsored research and continuing medical education. JAMA 2001; 286: 302-304.
14. Hansson L, Zanchetti A, Carruthers SG, Dahlof B, Elmfeldt D, Julius S, et al., for the HOT Study Group. Effects of intensive blood pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. Lancet 1998; 351:1755-1762.
15. Sandham JD, Hull RD, Brand RF, Knox L, Pineo GF, Doig CJ, et al., for the Canadian Critical Care Clinical Trials Group. A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients. N Engl J Med 2003; 348:5-14.
16. LIPID Study Group. Long-term effectiveness and safety of pravastatin in 9014 patients with coronary heart disease and average cholesterol concentrations: the LIPID trial follow-up. Lancet 2002; 359:1379-1387.
17. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002; 360:7-22.
18. Parisi AF, Folland ED, Hartigan P, on behalf of the Veterans Affairs ACME Investigators. A comparison of angioplasty with medical therapy in the treatment of single-vessel coronary artery disease. N Engl J Med 1992; 326:10-16.
19. Seppälä H, Nissinen A, Järvinen H, Huovinen S, Henrikson T, Herva E, et al. Resistance to erythromycin in group A streptococci. N Engl J Med 1992; 326:292-297.
20. Rafal ES, Griffiths CE, Ditre CM, Finkel LJ, Hamilton TA, Ellis CN, et al. Topical retinoin treatment for liver spots associated with photodamage. N Engl J Med 1992; 326: 368-374.
21. Barrett BJ, Parfrey PS, Vavasour HM, O'Dea F, Kent G, Stone E. A comparison of nonionic, low-osmolality radiocontrast agents with ionic, high-osmolality agents during cardiac catheterization. N Engl J Med 1992; 326:431-436.

22. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948; 2: 769-782.
23. Motulsky H. P-values, definition and common misinterpretations. In: Intuitive biostatistics. Ed Motulsky H, Oxford University Press, New York, 1995, pp96-97.

CHAPTER 11

RESEARCH DATA CLOSER TO EXPECTATION THAN COMPATIBLE WITH RANDOM SAMPLING

1. INTRODUCTION

Research data may be close to expectation. However, a difference from control of 0.000 is hardly compatible with random sampling. As it comes to well-balanced random sampling of representative experimental data, nature will be helpful to provide researchers with results close to perfection. However, because biological processes are full of variations, nature will never allow for 100% perfection. Statistical distributions can account for this lack of perfection in experimental data sampling, and provide exact probability levels of finding results close to expectation.

As an example, in a Mendelian experiment the expected ratio of yellow-peas / green-peas is 1 / 1. A highly representative random sample of n = 100 might consist of 50 yellow and 50 green peas. However, the larger the sample the smaller the chance of finding exactly fifty/fifty. The chance of exactly 5000 yellow / 5000 green peas or even the chance of a result very close to this result is, due to large variability in biological processes, almost certainly zero. In a sample of 10,000 peas, you might find 4997 yellow and 5003 green peas. What is the chance of finding a result this close to expectation? A simple chi-square test produces here a $p > 0.95$ of finding a result less close, which means a chance of $< (1-0.95)$, i.e., < 0.05 of finding a result this close or closer. Using the traditional 5% decision level, this would mean, that we have a strong argument that these data are not completely random. The example is actually based on some true historic facts, Mendel improved his data.[1]

Mendel's data were unblinded and unrandomized. Currently interventional data are obtained through randomized controlled trials. The phenomenon of data closer to expectation than compatible with random sampling is not considered anymore. But it is unknown whether it has actually disappeared. In the previous chapter the subject of extreme p-values as a result of research data closer to expectation than compatible with random sampling has been briefly addressed. The current chapter provides additional methods and examples in order to further emphasize the importance of this issue.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 137–143.

2. METHODS AND RESULTS

In order to assess this issue we defined data closer than random according to:

1.An observed p-value of > 95 %.
This literally means that we have > 95 % chance of finding a result less close expectation, and, consequently, < 5 % chance of finding a result this close or closer.

2.An observed p-value of < 0.0001.
Often in the study protocol a statistical power of 80% is agreed, corresponding with a p-value of approximately 0.01. The ultimate p-value may then be a bit larger or smaller. However a p-value of > 0.05 will be rarely observed, because current clinical trials are confirmational and, therefore, rarely negative. Also a p-value much smaller than 0.01 will be rarely observed, because it would indicate that the study is overpowered. If the p-values can be assumed to follow a normal distribution around p = 0.01, then we will have less than 5% chance of observing a p-value of < 0.0001.

3.An observed standard deviation (SD) < 50% the SD expected from prior population data.
From population data we can be pretty sure about SDs to be expected. E.g., the SDs of blood pressures are close to 10% of their means, meaning that for a mean systolic blood pressures of 150 mm Hg the expected SD is close to 15 mm Hg, for a mean diastolic blood pressure of 100 mm Hg the expected SD is close to 10 mm Hg. If such SDs can be assumed to follow a normal distribution, we will have < 5% chance of finding SDs < 7.5 and < 5 mm Hg respectively.

4.An observed standard deviation (SD) > 150% the SD expected from prior population data.
With SDs close to 10% of their means, we, likewise, will have <5% chance of finding SDs >150% the size of the SDs expected from population data.

We, then, searched randomized controlled trials of the 1999-2002 volumes of four journals accordingly. However, we decided to early terminate our search after observing respectively 7, 14, 8 and 2 primary endpoint results closer than random in a single random issue from the journals (Table 1). We have to conclude that the phenomenon of research data closer to expectation than compatible with random sampling has not at all disappeared. We assume that, like with the above Mendelian example, inappropriate data cleaning is a major factor responsible. We recommend that the statistical community develop guidelines for assessing appropriateness of data cleaning, and that journal editors require submitters of research papers to explain their results if they provide extremely high or low p-values or unexpectedly small or large SDs. Maybe, they should even consider, like the New England Journal of Medicine, not to publish p-values smaller than 0.001 anymore.

Table 1. Numbers of primary endpoint results closer to expectation than compatible with random sampling observed in a single issue from four journals

	p>0.95	p<0.0001	SD<50% of expected SD	>150% of expected SD
Cardiovascular Research 1999; 43: issue 1	1 (1)*	5 (1)	3 (2)	1 (1)
Current Therapeutic Research 2000; 61: issue 1	0 (0)	3 (1)	3 (1)	0 (0)
International Journal of Clinical Pharmacology and Therapeutics 2001; 39: issue 12	3 (2)	1 (1)	0 (0)	0 (0)
Journal of Hypertension 2002; 20: issue 10	3 (2)	5 (1)	2 (1)	1 (1)
Total	7 (5)	14 (4)	8 (4)	2 (2)

*Between brackets numbers of studies.

Evidence-based medicine is under pressure due to the conflicting results of recent trials producing different answers to similar questions.[2,3] Many causes are mentioned. As long as the possibility of inappropriate data cleaning has not been addressed, this very possibility cannot be excluded as potential cause of the obvious lack of homogeneity in current research.

3. DISCUSSION

In randomized controlled trials, prior to statistical analysis, the data are checked for outliers and erroneous data. Statistical tests are, traditionally, not very good at distinguishing between errors and outliers, but they should be able to point out main endpoint results closer to expectation than compatible with random sampling. In the current chapter we propose some criteria to assess main endpoint results for such purpose. One of the criteria proposed is a <5% probability to observe p-values of <0.0001 in studies planned at a power of 80%.[4] Kieser[5] takes issue with this proposed criterion, and states that, based on the two-sample normally distributed model of Hung[6], this probability should be much larger than 5 %. We used a different, and, in our view, more adequate model for assessment, based on the t-distribution and a usual two-sided type I error of 5%, rather than a one-sided type I error of 1%. We here take the opportunity to explain our assessment a little bit further and, particularly, to explain the arguments underlying it.

In statistics, a generally accepted concept is “the smaller the p-value, the better reliable the results”. This is not entirely true with current randomized controlled trials. First, randomized controlled trials are designed to test small differences. A randomized

controlled trial with major differences between old and new treatment is unethical because half of the patients have been given an inferior treatment. Second, they are designed to confirm prior evidence. For that purpose, their sample size is carefully calculated. Not only too small, but also too large a sample size is considered unethical and unscientific, because negative studies have to be repeated and a potentially inferior treatment should not be given to too many patients. Often in the study protocol a statistical power of 80% is agreed, corresponding with a p-value of approximately 0.01 (Figure 1). The ultimate p-value may then be a little bit larger or smaller. However, a p-value > 0.05 will be rarely observed, because most of the current clinical trials are confirmational, and, therefore, rarely negative. Also, a p-value much smaller than 0.01 will be rarely observed, because it would indicate that either the power assessment was inadequate (the study is overpowered) or data management was not completely adequate. With p = 0.0001 we have a peculiar situation. In this situation the actual data can not only reject the null-hypothesis (H_0, Figure 2) at p = 0.0001, but also the hypothesis of significantly better (H_1, Figure 2) at p = 0.05. This would mean that not only H_0 but also H_1 is untrue.

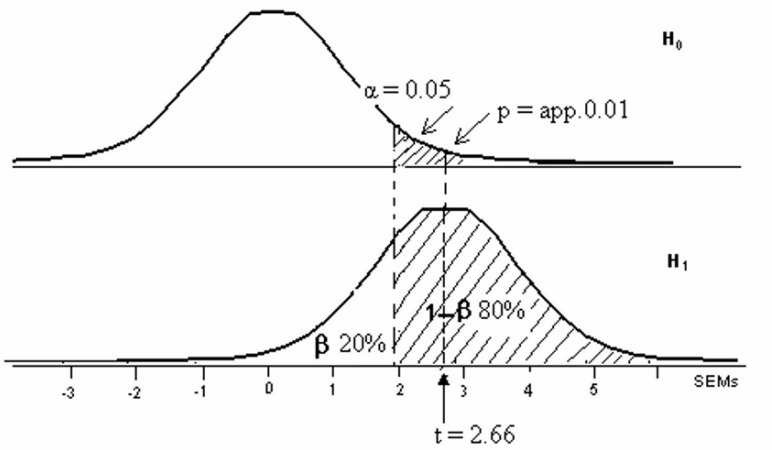


Figure 1. Null-hypothesis (H0) and alternative hypothesis (H1) of an example of experimental data with sample size n = 60 and mean = 2.66, and a t-distributed frequency distribution. The null-hypothesis is rejected with a p-value of approximately 0.01 and a statistical power (=1- β) of 80%.

α = type I error = 5%; β = type II error = 20%; app.= approximately; SEM = standard error of the mean.

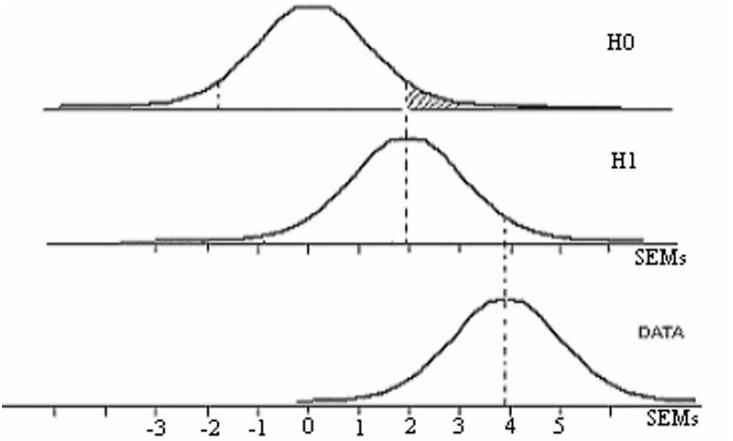


Figure 2. Null-hypothesis (H0), hypothesis of significantly better (H1), and actual data distribution (DATA) of an example of experimental data with n = 120 and mean = 3.90 SEMs and a t-distributed frequency distribution. The actual data can not only reject H0 (t = 3.90, p = 0.0001), but also H1 (t = 1.95, p = 0.05). This would mean that not only H0 but also H1 is untrue.

SEM = standard error of the mean.

Table 2 gives an overview of five published studies with main endpoint p-values < 0.0001. All of these studies were published in the first 6 issues of the 1992 volume of the New England Journal of Medicine. It is remarkable that so many overpowered studies were published within 6 subsequent months of a single volume, while the same journal published not any study with p-values below 0.001 in the past 4 years' full volumes. We do not know why, but this may be due to the journal's policy not to accept studies with very low p-values anymore. In contrast, many other journals including the Lancet, Circulation, BMJ (British Medical Journal), abound with extremely low p-values. We should add that, while preparing this chapter, we noticed that, in the past two months, also JAMA (Journal American Medical Association) did not publish p-values below 0.001 anymore. It is obvious, however, that most of the other journals still believe in the concept "the lower the p-value, the better reliable the research". The concept may still be true for observational studies. However, in conformational randomized controlled trials, p-values as low as 0.0001 do not adequately confirm prior hypotheses anymore, and have to be checked for adequacy of data management.

Table 2. Study data with p-values as low as <0.0001, published in the first 6 issues of the 1992 volume of the New England Journal Medicine. In the past 4 years p-values smaller than p<0.001 were never published in this journal

	Result	Sample size requirement	alpha-level	p-values
1. Ref. 7	+0.5 vs +2.1%	yes	0.05	< 0.0001
2. Ref. 7	–2.8 vs +1.8 %	yes	0.05	< 0.0001
3. Ref. 8	11 vs 19 %	no	0.05	< 0.0001
4. Ref. 9	r = - 0.53	no	0.05	< 0.0001
5. Ref. 10	213 vs 69	no	0.05	< 0.0001

1.Duration exercise in patients after medical therapy vs percutaneous coronary angioplasty. 2. Maximal double product (systolic blood pressure times heart rate) during exercise in patients after medical treatment vs percutaneous coronary angioplasty. 3. Erythromycin resistance throat swabs vs pus samples. 4. Correlation between reduction of epidermal pigmentation during treatment and baseline amount of pigmentation. 5. Adverse reactions of high vs non-high osmolality agents during cardiac catheterization. Alpha = type I error, vs= versus.

4. CONCLUSIONS

The following results may be closer to expectation than compatible with random.
1.An observed p-value of > 0.95.
2.An observed p-value of < 0.0001.
3.An observed standard deviation (SD) < 50% the SD expected from prior population data.
4.An observed standard deviation (SD) > 150% the SD expected from prior population data.
Additional assessments to identify data at risk of unrandomness will be reviewed in chapter 24.

5. REFERENCES

1. Cleophas TJ, Cleophas GM. Sponsored research and continuing medical education. JAMA 2001; 286: 302-4.
2. Julius S. The ALLHAT study: if you believe in evidence-based medicine, stick to it. J Hypertens 2003; 21: 453-4.
3. Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. Int J Clin Pharmacol Ther 2003; 41 51-6.
4. Cleophas TJ. Research data closer to expectation than compatible with random sampling I. Stat Med 2004; 23: 1015-7.
5. Kieser M, Cleophas TJ. Research data closer to expectation than compatible with random sampling II. Stat Med 2005; 24: 321-3.

6. Hung HMJ, O'Neill RT, Bauer P, Köhne K. The behavior of the p-value when the alternative hypothesis is true. Biometrics 1997; 53: 11-22.
7. Parisi AF, Folland ED, Hartigan P, on behalf of the Veterans Affairs ACME Investigators. A comparison of angioplasty with medical therapy in the treatment of single-vessel coronay artery disease. N Engl J Med 1992; 326: 10-6.
8. Seppälä H, Nissinen A, Järvinen H, Huovinen S, Henrikson T, Herva E, et al. Resistance to erythromycin in group A streptococci. N Engl J Med 1992; 326: 292-7.
9. Rafal ES, Griffiths CE, Ditre CM, Finkel LJ, Hamilton TA, Ellis CN, et al. Topical retinoin treatment for liver spots associated with photodamage. N Engl J Med 1992; 326: 368-74.
10. Barrett BJ, Parfrey PS, Vavasour HM, O'Dea F, Kent G, Stone E. A comparison of nonionic, low-osmolality radiocontrast agents with ionic, high-osmolality agents during cardiac catheterization. N Engl J Med 1992; 326: 431-6.

CHAPTER 12

STATISTICAL TABLES FOR TESTING DATA CLOSER TO EXPECTATION THAN COMPATIBLE WITH RANDOM SAMPLING

1. INTRODUCTION

A p-value <0.05 is generally used as a cut-off level to indicate a significant difference from what we expect. A p-value of > 0.05, then, indicates no significant difference. The larger the p-value the smaller the chance of a difference. A p-value of 1.00 means 0 % chance of a difference, while a p-value of 0.95 means a chance of difference close to 0. A p-value of > 0.95 literally means that we have > 95 per cent chance of finding a result less close to expectation, which means a chance of < (1-0.95) , i.e., < 0.05 of finding a result this close or closer. Using the traditional 5 per cent decision level, this would mean, that we have a strong argument that such data are not completely random. The example from the previous chapter is used once more. In a Mendelian experiment the expected ratio of yellow-peas/-green peas is 1/1. A highly representative random sample of n = 100 might consist of 50 yellow and 50 green peas. However, the larger the sample the smaller the chance of finding exactly fifty / fifty. The chance of exactly 5000 yellow / 5000 green peas or even the chance of a result very close to this result is, due to large variability in biological processes, almost certainly zero. In a sample of 10,000 peas, you might find 4997 yellow and 5003 green peas. What is the chance of finding a result this close to expectation? A chi-square test produces here a p > 0.95 of finding a result less close, and consequently, < 0.05 of finding a result this close or closer. Using the 5% decision level, this would mean, that we have a strong argument that these data are not completely random. The example is actually based on some true historic facts, Mendel improved his data.[1]

Some readers might be confused by the assertion that a p-value >0.95 implies that the data are closer to expectation than compatible with random sampling. A large p-value, indeed, generally, means that the data behave very similar to that which one would expect under the null hypothesis. Yet, a p-value >0.95 will be rarely observed, because not only data but also *mean outcomes of data-samples* follow (normal or chi-square) frequency distributions. Figure 1 displays a chi-square-

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 145–153.

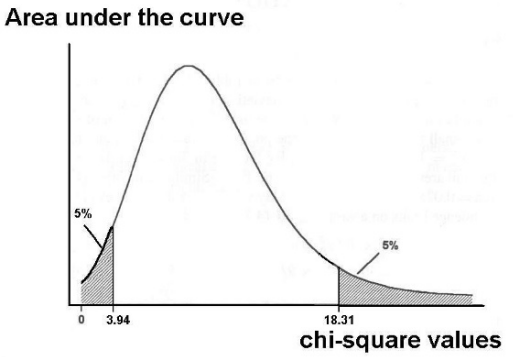


Figure 1. Null hypothesis of chi-square distributed samples with 10 degrees of freedom. The area under the curve for chi-square values larger than 18.31 or smaller than 3.94 is < 5%. This means we have <5% chance to find a variability thát large or thát small, and so, we are entitled to reject the null hypothesis in either case.

distributed null hypothesis curve (10 degrees of freedom). On the x-axis we have the so-called chi-square values which can be interpreted as estimates of variabilities of studies that have 10 degrees of freedom. On the y-axis we have p-values (= areas under the curve). The curve presents the collection of all of the variabilities one can expect. The area under the curve for chi-square values larger than 18.31 or smaller than 3.94 is < 5%. This means we have <5% chance to find a variability thát large or thát small, and so, we are entitled to reject the null hypothesis in either case. The left end of the chi-square curve, although routinely used for testing appropriateness of data distributions, is little used for the above purpose so far.

In a recent search of randomized trials published in four journals we found main-endpoint results with p-values > 95 percent in every single issue of the journals.[2] We assumed that inappropriate data cleaning was a major factor responsible. In clinical research the appropriateness of data cleaning is rarely assessed. The current paper was written to facilitate the assessment of this issue. We present tables of unusually large p-values to assess data closer to expectation than compatible with random sampling. We also give examples showing how to calculate such p-values from published data yourself, and examples to simulate real practice. The paper tries to address a phenomenon, rather than accuse research groups. Therefore, only simulated examples are given.

2. STATISTICAL TABLES OF UNUSUALLY HIGH P-VALUES

Statistical tests estimate the probability that a difference in the data is true rather than due to chance, otherwise called random. For that purpose they make use of test-statistics:

test-statistic	test
t-value	for the t-test
chi-square	for the chi-square test
F-value	for the F-test.

The t-statistic (t-value) is used for the assessment of the means of continuous data, odds ratios and regression coefficients. The chi-square statistic (chi-square-value) is used for the analysis of proportional data, and survival data. The f-statistic (f-value) is used for comparing continuous data from more than two groups or more than two observations in one person, and for additional purposes such as testing correlation coefficients. The Tables 1-3 give overviews of the unusual sizes these test-statistics and their corresponding p-values can adopt if data are closer to expectation than compatible with random sampling.

3. HOW TO CALCULATE THE P-VALUES YOURSELF

t-test

In a parallel-group study two cholesterol reducing drugs are assessed.
Group1 (n=50), mean result 3.42 mmol/l, standard error of the mean (SEM) 0.06 mmol / l.

Table 1. t-values with unusually high p-values

	p-value (two sided)			
	0.999	0.99	0.95	0.90
degrees freedom				
1	0.0015	0.154	0.0770	0.1580
2	0.0014	0.141	0.0707	0.1419
3	0.0014	0.136	0.0681	0.1366
4	0.0013	0.0133	0.0667	0.1338
5	0.0013	0.0132	0.0659	0.1322
6	0.0013	0.0132	0.0654	0.1311
7	0.0013	0.0130	0.0650	0.1303
8	0.0013	0.0129	0.0647	0.1297
9	0.0013	0.0129	0.0647	0.1293
10	0.0013	0.0129	0.0643	0.1289

15	0.0013	0.0127	0.0638	0.1278
20	0.0013	0.0127	0.0635	0.1273
30	0.0013	0.0126	0.0632	0.1267
40	0.0013	0.0126	0.0631	0.1265
50	0.0013	0.0126	0.0630	0.1263
60	0.0013	0.0126	0.0630	0.1262
70	0.0013	0.0126	0.0629	0.1261
80	0.0013	0.0126	0.0629	0.1261
100	0.0013	0.0126	0.0629	0.1261
∞	0.0013	0.0126	0.0629	0.1261

Group2 (n = 50), mean result 3.38 mmol/l, SEM 0.06 mmol / l.
Difference between results 0.04 mmol /l, pooled SEM = $\sqrt{SEM_1^2 + SEM_2^2} = 0.848$.
T-value= 0.04 / 0.848= 0.0472 for (50+50) -2 = 98 degrees of freedom. T-value is <0.0619, and according to the t-table (Table 1) the p-value is thus > 0.95. These data are closer to expectation than compatible with random.

chi-square test
The underneath example is given in the form of a 2 x 2 contingency table which follows a chi-square distribution with 1 degree of freedom.

Pea phenotype		P	p
	R	PR 27 (a)	pR 271(b)
	r	Pr 9 (c)	pr 92 (d)

The appropriate chi-square value according to:

$$\frac{(ad-bc)^2 (a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = 0.00205$$

The chi-square table (Table 2) gives in four columns various chi-square values corresponding to the p-values given in the upper row. The left hand column adjusts

Table 2. Chi-square values with unusually high p-values

	p-value(two sided)			
	0.999	0.99	0.95	0.90
degrees freedom				
1	0.0000016	0.0016	0.0039	0.016
2	0.0020	0.020	0.10	0.21
3	0.024	0.12	0.35	0.58
4	0.091	0.30	0.71	1.06
5	0.21	0.55	1.15	1.61

6	0.38	0.87	1.64	2.20
7	0.60	1.24	2.17	2.83
8	0.86	1.65	2.73	3.49
9	1.15	2.09	3.33	4.17
10	1.48	2.56	3.94	4.87
15	3.48	5.23	7.26	8.55
20	5.92	8.26	10.85	12.44
25	8.66	11.52	14.61	16.47
30	11.58	14.95	18.49	20.60
35	14.68	18.51	22.46	24.80
40	17.93	22.16	26.51	29.05
50	24.68	29.71	34.76	37.69
60	31.73	37.49	43.19	46.46
70	39.02	45.44	45.74	55.33
80	46.49	53.54	60.39	64.28
100	61.92	70.07	77.93	82.36

for degrees of freedom. The above result of 0.00205 is on the left side of the critical chi-square value of 0.0039, and, so, we can reject the possibility that our data are so close to each other by chance even if there were no difference in the data. We may worry that these data are not randomly sampled.

F-test

The effect of 3 compounds improving Hemoglobin levels is assessed in three parallel groups.

Group	n patients	mean(mmol/l)	SD(mmol/l)
1	16	10.6300	1.2840
2	16	10.6200	1.2800
3	16	10.6250	1.2820

Grand mean = (mean 1 + 2 +3)/3 = 10.6250

$SS_{between\ groups}$=16 (10.6300-10.6250) ^{2} + 16(10.6200 – 10.6250)^2 + 16(10.6250-10.6250)^2

= 0.000625 for 3 groups meaning 3-1=2 degrees freedom

$MS_{between\ groups}$ = 0.000625 / 2 = 0.0003125

$SS_{within\ groups}$ = 15 x 1.2840^2 + 15 x 1.2800^2 +15 x 1.2820^2

= 24.730 + 14.746 + 23.009 = 62.485 for 15+9+14= 38 degrees of freedom

$MS_{within\ groups}$ = 62.485 / 38= 1.644

F = $MS_{between\ groups}$ / $MS_{within\ groups}$ = 0.0003125 / 1.644 = 0.00019

According to the F-table (Table 3) for 2 and 38 degrees of freedom an F-value >0.051 means that p > 95 per cent. The differences between the parallel groups are

Table 3. F-values with unusually high p-values

degrees of freedom of the numerator															
1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	1000
degrees of freedom denominator															
1	0.0062	0.054	0.099	0.13	0.15	0.17	0.18	0.19	0.20	0.20	0.22	0.23	0.23	0.25	0.25	0.26
2	0.0050	0.053	0.10	0.14	0.17	0.19	0.21	0.22	0.23	0.24	0.27	0.29	0.3 0	0.31	0.32	0.33
3	0.0046	0.052	0.11	0.15	0.18	0.2 1	0.23	0.25	0.26	0.27	0.30	0.32	0.34	0.36	0.37	0.38
4	0.0045	0.052	0.11	0.16	0.19	0.22	0.24	0.26	0.28	0.29	0.33	0.35	0.37	0.39	0.41	0.42
5	0.0043	0.052	0.11	0.16	0.20	0.23	0.25	0.27	0.29	0.30	0.34	0.37	0.40	0.42	0.43	0.45
6	0.0043	0.052	0.11	0.16	0.20	0.23	0.26	0.28	0.30	0.31	0.36	0.38	0.41	0.44	0.46	0.47
7	0.0042	0.052	0.11	0.16	0.20	0.24	0.26	0.29	0.30	0.32	0.37	0.40	0.43	0.45	0.48	0.50
8	0.0042	0.052	0.11	0.17	0.21	0.24	0.27	0.29	0.31	0.33	0.39	0.41	0.44	0.47	0.49	0.51
9	0.0041	0.052	0.11	0.17	0.21	0.24	0.27	0.29	0.31	0.33	0.39	0.42	0.45	0.48	0.51	0.53
10	0.0041	0.052	0.11	0.17	0.21	0.24	0.27	0.30	0.32	0.34	0.39	0.43	0.46	0.49	0.52	0.54
15	0.0041	0.052	0.11	0.17	0.22	0.25	0.28	0.31	0.33	0.35	0.42	0.45	0.50	0.53	0.56	0.60
20	0.0040	0.052	0.11	0.17	0.22	0.26	0.29	0.32	0.34	0.36	0.43	0.47	0.52	0.56	0.60	0.63
30	0.0040	0.051	0.12	0.17	0.22	0.26	0.30	0.32	0.35	0.37	0.44	0.49	0.54	0.59	0.64	0.68
50	0.0040	0.051	0.12	0.17	0.23	0.27	0.30	0.33	0.36	0.38	0.46	0.51	0.57	0.63	0.67	0.74
100	0.0040	0.051	0.12	0.18	0.23	0.27	0.31	0.34	0.36	0.39	0.47	0.52	0.59	0.66	0.72	0.79

closer to zero than compatible with random sampling. We have a strong argument to believe that they are not completely random.

4. ADDITIONAL EXAMPLES SIMULATING REAL PRACTICE, MULTIPLE COMPARISONS

The issue of p-values > 0.95 being not random is, of course, less true for studies testing multiple measurements. If you test many times, you are apt to find extreme p-values, either high or low, once in a while purely by chance. If the chance of finding an extreme p-value with a single test is equal to 0.05, then, according to the Bonferroni inequality, this chance increases to $1\text{-}0.95^k$ with k tests. However, the chance of finding multiple extreme p-values in this situation remains very small. E.g., the chance of finding k extreme p-values with k tests is equal to 0.05^k. If k= 5, then this chance = <0.000000. We should add that multiple comparisons/tests will not be independent in most cases. Therefore, the chance of finding extreme p-values will not be as dramatically small as implied by the above formula, but even with dependencies, this chance will soon get much smaller than the traditional 0.05, and unrandomness of data has to be accounted.

Table 4 gives examples of multiple comparisons/tests as commonly included in the reports of clinical drug trials/research. It show a remarkable similarity of (1) patient characteristics between two treatment groups, and (2) pharmacokinetic data between a brand name drug and its generic copy, (3) the virtual absence of time or

carryover effect in a crossover study, and (4) virtually no difference in side effects between treatment and a placebo. All of these examples include multiple comparisons/tests in a single population, and these comparisons/tests can, therefore, not be expected to be independent of one another. The chance that these

Table 4. Examples of multiple comparisons/tests as commonly included in the reports of clinical drug trials / research

Example 1 Patient characteristics of a randomized controlled trial (sds=standard deviations)

	treatment 1 (n=5000)	**treatment 2 (n=5000)**	**p - value**
females n(%)	979 (19.58)	974 (19.48)	>0.95
age <60 years n(%)	1882 (37.64)	1877 (37.54)	>0.95
white n (%)	4889 (97.78)	4887 (97.74)	>0.99
smokers n(%)	1716 (34.32)	1712 (34.32)	>0.95
mean alcohol consumption units(sds)	8.1(11.3)	8.0 (11.4)	>0.95
mean systolic blood pressure (mmHg, sds)	164.2 (17.8)	164.2 (17.8)	1.00
mean diastolic blood pressure (mmHg, sds)	95.0 (10.3)	95.0 (10.3)	1.00
mean body mass index (kg/m^2 , sds)	28.6 (4.7)	28.7 (4.6)	1.00
mean total cholesterol (mmol/l, sds)	5.5 (0.8)	5.5 (0.8)	1.00
mean triglycerides (mmol/l,sds)	1.74 (0.91)	1.73 (0.90)	>0.99
mean glucose (mmol/l, sds)	6.2 (2.1)	6.2 (2.1)	1.00

Example 2 Pharmacokinetic parameters

	brand-name drug (n=8)	**generic copy (n=7)**	**p - value**
mean clearance (ml/hr, sds)	158 (15)	158 (15)	1.00
mean bio-availibility(%, sds)	75 (7)	74 (7)	>0.95
mean volume of distribution (ml)	9300 (910)	9296 (915)	>0.99
mean elimination half life (hr, sds)	41 (4)	40 (4)	>0.95
mean area under curve (microg,hr/ml)	547 (54)	543 (53)	>0.95

Example 3 Crossover study tested for treatment, carryover and time effects

	period 1	**period 2**
	mean temperature (^{0}C, sd)	
group1	treatment 1	treatment 2
	result 20.23 (4.12) *a*	result 24.12 (4.21) *b*
group2	treatment 2	treatment 1
	result 24.11 (4.20) *c*	result 20.22 (4.12) *d*

treatment effect $a+d$ vs $b+c$ $p < 0.001$

carryover effect $a+c$ vs $b+d$ $p > 0.95$
time effect $a+b$ vs $c+d$ $p > 0.95$

Continued Table 4.

Example 4 Side effects of a parallel-group clinical trial

	active treatment (n=500)	**placebo (n=500)**	**p - value**
nasal congestion (yes)	240	241	>0.99
urine incontinence	44	45	>0.95
impotence	50	52	>0.95
depression	32	31	>0.95
fatigue	99	98	>0.95
palpitations	50	50	1.00
dizziness	121	123	>0.95
sleepiness	76	80	>0.95

tables are unrandom is, therefore, not as small as implied by the above formula, but it is, certainly, smaller than 5% for each of the examples given.

5. DISCUSSION

Main-endpoint results producing large p-values may not be entirely random. This issue has received little attention from the scientific community so far. Also statistical tables covering them are not in the statistical literature. Fortunately, current statistical software generally provides exact p-values. But, then, investigators, however excited to report how nicely their results match their prior expectations, are often reluctant to report the exact p-values, and confine themselves to the notion NS (not significant). The statistical tables as published in the current paper can be adequately used to test, a posteriori, such data. Whenever, p-values are >0.95, we have a strong argument that the data are not entirely random, and that they be interpreted with caution.

The current chapter is only a preliminary effort to assess randomness of clinical trial data. Other methods could include the more extensive use of population data for comparison, data transformations and non-parametric tests. We recommend that the scientific community develop guidelines for standard assessment of this issue. This is important to the body of evidence-based medicine, currently under pressure due to the conflicting results of recent trials producing different answers to similar questions.[3,4] Many causes are mentioned. As long as the possibility of

inappropriate data cleaning has not been addressed, this very possibility can not be excluded as potential cause of the obvious lack of homogeneity in current research.

6. CONCLUSIONS

A p-value of > 0.95 literally means that we have > 95 per cent chance of finding a result less close to expectation, and, consequently, <5 per cent chance of finding a result this close or closer. Using the traditional 5 per cent decision level, this would mean, that we have a strong argument that such data are not completely random. The objective of this chapter was to facilitate the assessment of this issue. T-, chi-square-, and f-tables of unusually large p-values are given to calculate a posteriori p-values of study results closely matching their prior expectations. Simulated examples are given. Clinical trial data producing large p-values may not be completely random. The current chapter is a preliminary effort to assess randomness of clinical trial data.

7. REFERENCES

1. Cleophas TJ, Cleophas GM. Sponsored research and continuing medical education. J Am Med Assoc 2001; 286: 302-4.
2. Cleophas TJ. Research data closer to expectation than compatible with random sampling. Stat Med 2004; 23: 1015-7.
3. Julius S. The ALLHAT study: if you believe in evidence-based medicine, stick to it. J Hypertens 2003; 21: 453-4.
4. Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. Int J Clin Pharmacol Ther 2003; 41: 51-6.

CHAPTER 13

PRINCIPLES OF LINEAR REGRESSION

1. INTRODUCTION

In the past chapters we discussed different statistical methods to test statistically experimental data from clinical trials. We did not emphasize correlation and regression analysis. The point is that correlation and regression analysis test correlations, rather than causal relationships. Two samples may be strongly correlated e.g., two different diagnostic tests for assessment of the same phenomenon. This does, however, not mean that one diagnostic test causes the other. In testing the data from clinical trials we are mainly interested in causal relationships. When such assessments were statistically analyzed through correlation analyses mainly, we would probably be less convinced of a causal relationship than we are while using prospective hypothesis testing. So, this is the main reason we so far did not address correlation testing extensively. With epidemiological observational research things are essentially different: data are obtained from the observation of populations or the retrospective observation of patients selected because of a particular condition or illness. Conclusions are limited to the establishment of relationships, causal or not. We currently believe that relationships in medical research between a factor and an outcome can only be proven to be causal if the factor is introduced, and, subsequently, gives rise to the outcome. We are more convinced when such is tested in the form of a controlled clinical trial. A problem with multiple regression and logistic regression analysis as method for analyzing multiple samples in clinical trials is closely related to this point. There is always an air of uncertainty about such regression data. Interventional trials usually use hypothesis-testing and 95 % confidence intervals (CIs) of the data to describe and analyze data. They use multiple regression for secondary analyses, thus enhancing the substance of the research, and making the readership more willing to read the report, rather than proving the primary endpoints. Regression analysis may not be so important to randomized clinical trials, it is important to one particular study design, the crossover study, where every patient is given in random order test-treatment and standard treatment (or placebo). Figure 1 gives three hypothesized examples of crossover trials. It can be observed from the plots that in the left and right graph there seems to be a linear relationship between treatment one and two. The strength of relationship is expressed as r (=correlation coefficient) which varies between -1 and +1. The strongest association is given by either -1 or +1 (all data exactly on the line), the weakest association 0 (all data are parallel either to x-axis or to y-axis, or half one direction, half the other. A positive correlation in a crossover study is observed if

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 155–173.

two drugs from one class are compared. The patients responding well to the first drug are more likely to respond well to the second. In contrast, in crossover studies comparing drugs from different classes a negative correlation may be observed: patients not responding well to one class are more likely to respond well to the other.

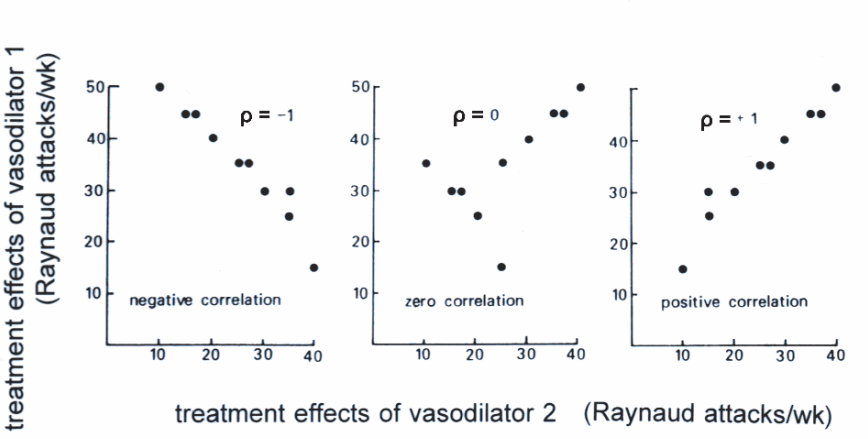


Figure 1. Example of 3 crossover studies of two treatments in patients with Raynaud's phenomenon. the (Pearson's) correlation coefficient ρ varies between -1 and +1.

2. MORE ON PAIRED OBSERVATIONS

Table 1 gives the real data of a crossover study comparing a new laxative versus a standard laxative, bisacodyl. Days with stool are used as primary endpoint. The table shows that the new drug is more efficacious than bisacodyl, but the figure (Figure 2) shows something else: there is a positive correlation between the two treatments: those responding well to bisacodyl are more likely to respond well to the novel laxative.

Table 1. Example of a crossover trial comparing efficacy of a new laxative versus bisacodyl

patient no.	new treatment (y-variables) (days with stool)	bisacodyl (x-variables) (days of stool)
1	24	8
2	30	13
3	25	15
4	35	10
5	39	9
6	30	10
7	27	8
8	14	5
9	39	13
10	42	15
11	41	11
12	38	11
13	39	12
14	37	10
15	47	18
16	30	13
17	36	12
18	12	4
19	26	10
20	20	8
21	43	16
22	31	15
23	40	14
24	31	7
25	36	12
26	21	6
27	44	19
28	11	5
29	27	8
30	24	9
31	40	15
32	32	7
33	10	6
34	37	14
35	19	7

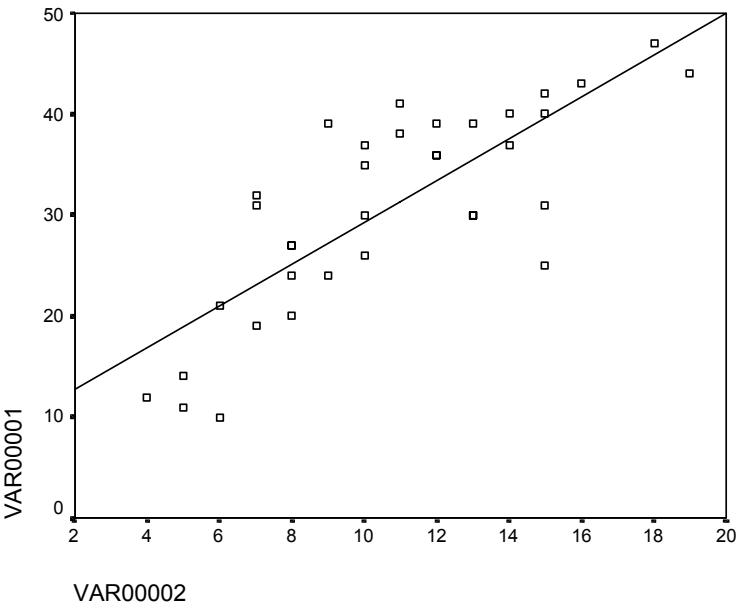


Figure 2. Scatterplot of data from Table 1 with regression line.

A regression line can be calculated from the data according to the equation

$$y=a+bx$$

The line drawn from this linear function provides the best fit for the data given, where y = socalled dependent, and x = independent variable, b = regression coefficient.

a and b from the equation y=a+bx can be calculated.

$$b = \text{regression coefficient} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$a = \text{intercept} = \bar{y} - b\bar{x}$$

r = correlation coefficient = another important determinant and looks a lot like b.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

r = measure for the strength of association between y and x-data. The stronger the association, the better y predicts x.

3. USING STATISTICAL SOFTWARE FOR SIMPLE LINEAR REGRESSION

Regression analysis without software is laborious. We may use a computer program, e.g., **SPSS Statistical Software**, to do the job for us. We command our software: **Statistics; Regression; Linear**. Excel files can be entered simply cutting and pasting.

The software calculates the values b and a and r so as to minimize the sum of the squared vertical distances of the points from the line (least squares fit). **SPSS 8 for windows 99** provides us with three tables (Table 2) : **(1) Model Summary, (2) ANOVA, (3) coefficients.**

Table 2.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.794[a]	.630	.618	6.1590

a. Predictors: (Constant), VAR00002

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2128.393	1	2128.393	56.110	.000[a]
	Residual	1251.779	33	37.933		
	Total	3380.171	34			

a. Predictors: (Constant), VAR00002

b. Dependent Variable: VAR00001

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.647	3.132		2.761	.009
	VAR00002	2.065	.276	.794	7.491	.000

a. Dependent Variable: VAR00001

(1) Model Summary gives information on correlation coefficient (r) and its square (r^2) the coefficient of determination. A coefficient of determination of 0.63 indicates that 63% of the variation in the y variable is explained by variation in the x variable. The better the effect of bisacodyl the better the novel laxative is going to work. Adjusted r square is important for small samples only while std error of the estimate tells us something about the residue (variance not explained by the regression) and is equal to the square root of the Residual Mean Square.

At this point it is important to consider the following. Before doing any regression analysis we have to make the assumptions that our y - data are normally distributed and that variances in y-variable do not show a lot of difference, otherwise called heteroscedasticity (heteroscedasticity literally means “different standard deviations (SDs)”).

White's Test is a simple method to check for this. Chi-square table is used for that purpose.

if $n r^2 < \chi^2 (n)$ we don't have to worry about heteroscedasticity.

n= sample size
r= correlation coefficient
$\chi^2 (n)$ = the value for n degrees of freedom.
In our example 35 (0.630) = 22.05 while $\chi^2 (35) = 56.70$ (no heteroscedasticity)

(2) ANOVA (analysis of variance) shows how the paired data can be assessed in the form of analysis of variance. Variations are expressed as sums of squares. The total variation in the regression is divided into sum of squares (SS) regression, or variances explained by the regression, and SS residual, variances unexplained by the regression.

$$r^2 = \frac{(\sum (x - \overline{x})(y - \overline{y}))^2}{\sum (x - \overline{x})^2 \sum (y - \overline{y})^2} = \frac{SP^2 x \cdot y}{SSx \cdot SSy}$$

where SP = sum of products x . y

SS regression = SP^2 xy / SSx = 2128.393
SS total = SS y
SS regression/ SS total= 2128.393/ SS total = 0.63 (= r square (Model Summary))
As explained above this means that 63 % of the vraition in the y-varaible is explained buy the variation in the x-variable. This interpretation may be hard to understand, but it is helpful to imagine:
$r^2 = 0$ indicating no correlation at all,
r^2= 1.00 indicating 100% correlation, each y-datum is exactly on the line, $r^2 = 0.50$ indicating 50% certainty about a corresponding y- value if we know the x-value.
The strength of association of x- and y-values is dependent not only on the magnitude of the r^2 – value, but, in addition, on the sample size. For example, if we have a sample of n = 3 exactly on the line, then no accurate predictions can be made. However, if n = 100, then we are more convinced of the accuracy of the line as a predictor of y-values from given x-values. Therefore, in addition to calculating the magnitude of the r^2 –value, we have to include the sample size in our statistical work-up. For that purpose ANOVA is used. It tests whether r^2 is significantly larger than 0.00. The table shows that, indeed $p < 0.000$, and that the we, thus have a significant highly significant association between the x- and y-variables.
SPSS uses R (upper case), other software uses r (lower case) for expressing the correlation coefficient.

(3) Coefficients shows the regression equation. The intercept is named "(constant)" and IS given under B = 8.647. The b-value in the linear regression equation is 2.065.

The regression equation is thus as follows.

y= 8.647 + 2.065 . x

new laxative= 8.647 + 2.065 . bisacodyl

In addition to unstandardized coefficients, A standardized coefficients are given. For that purpose SSy is defined to be 1. Then, r = b. Instead of testing that is significantly larger than 0.00, we can now test that b is significantly larger than 0.000, and use for that purpose the t-test. The meaning of the two tests is very similar, and so is their result. The t-value of 7.491 = √ F= $\sqrt{56.110}$. This t-value is, obviously, equal to the square root of the F-value from the ANOVA-test.

4. MULTIPLE LINEAR REGRESSION

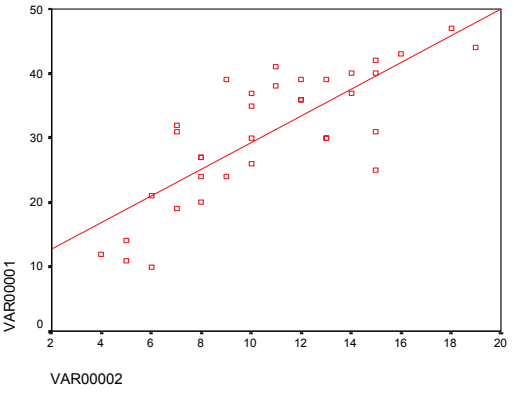


Figure 3. Scatterplot of data from Table 1 with regression line.

Linear regression would have never been so popular if only the inclusion of a single x-variable had been possible. We will now assess models with more than a single x-variable.

Obviously, there is a significant positive correlation between the x- and y-values in Figure 3 (the above laxative-study). Maybe, there is also a positive correlation between the new laxative and patient age. If so, then the new laxative might be better, e.g.,

(1) the better the bisacodyl,

(2) the older the patient.

In this case we have, thus, 3 observations in 1 person

(1) efficacy datum new laxative

(2) efficacy datum bisacodyl

(3) age.

In order to test possible correlations, we can define variables as follows

y variable presents new laxative data
x_1 variable bisacodyl data
x_2 variable age data.

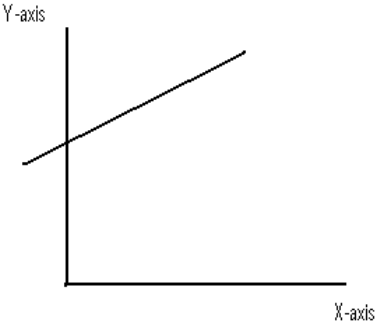


Figure 4. Linear regression model gives best predictable y-value for the x-value given.

Linear regression uses formula y=a + bx, where the y-variable = new laxative data, the x-variable = bisacodyl data. E.g., if we fill out

x-value =0 => then formula turns into y = a
x-value =1 => " " " " y = a + b
x-value =2 => " " " " y = a + 2b

For each x-value the formula produces the best predictable y-value, all y-values constitute a line, the regression line (Figure 4) which can be interpreted as the *best fit* line for data (the line with shortest distances from the y-values).

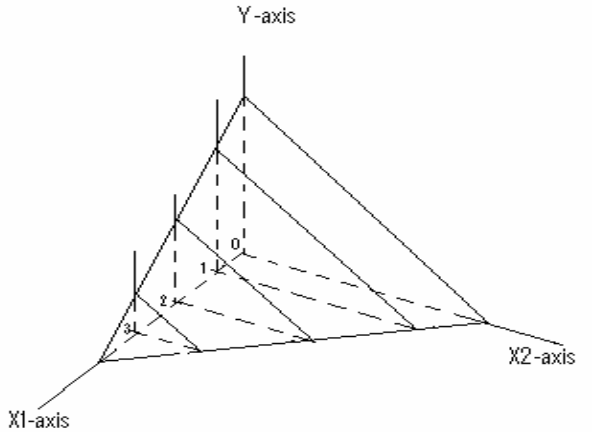


Figure 5. Three axes model to illustrate multiple linear regression model with two x-variables.

For multiple regression with 3 variables the regression formula $y = a + b_1 x_1 + b_2 x_2$ is being used. In order to visualize the model used, we can apply a 3-axes-model with y-axis, x_1 -axis and x_2 –axis (Figure 5). If we fill out

$x_1 = 0$, then the formula turns into $y = a + b_2 x_2$
$x_1 = 1$, « « « « « $y = a + b_1 + b_2 x_2$
$x_1 = 2$ « « « « « $y = a + 2b_1 + b_2 x_2$
$x_1 = 3$ « « « « « $y = a + 3b_1 + b_2 x_2$.

Each x_1 –value has its own regression line, all of the regression-lines constitute a regression plane which is interpreted as the best fit plane for the data (the plane with the shortest distances to the y-values).

5. MULTIPLE LINEAR REGRESSION, EXAMPLE

We may be interested to know if age is an independent contributor to the effect of the new laxative. For that purpose a simple regression equation has to be extended as follows

$$y = a + b_1 x_1 + b_2 x_2$$

b_i are called partial regression coefficients. Just like simple linear regression, multiple linear regression can give us the best fit for the data given. The calculations of a, b_1 and b_2 are given underneath.

$$\Sigma y = n a + b_1 \Sigma x_1 + b_2 \Sigma x_2$$
$$\Sigma x_1 y = a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x^2$$
$$\Sigma x_2 y = a \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

r between x_1 , x_2 en y calculate from the equation

$$R = \sqrt{(b_1 r x_1 + b_2 r x_2)}$$

The calculations are hard without a computer. Also, it is hard to display the correlations in a figure. Table 3 gives the data from table 1 extended by the variable age.

Table 3. Example of a crossover trial comparing efficacy of a new laxative versus bisacodyl

patient no.	new treatment y - variables (days with stool)	bisacodyl x_1 -variables (days with stool)	age x_2 -variables (years)
1	24	8	25
2	30	13	30

3	25	15	25
4	35	10	31
5	39	9	36
6	30	10	33
7	27	8	22
8	14	5	18
9	39	13	14
10	42	15	30
11	41	11	36
12	38	11	30
13	39	12	27
14	37	10	38
15	47	18	40
16	30	13	31
17	36	12	25
18	12	4	24
19	26	10	27
20	20	8	20
21	43	16	35
22	31	15	29
23	40	14	32
24	31	7	30
25	36	12	40
26	21	6	31
27	44	19	41
28	11	5	26
29	27	8	24
30	24	9	30
31	40	15	20
32	32	7	31
33	10	6	29
34	37	14	43
35	19	7	30

The Table 3 shows too many data to allow any conclusions. We use for assessment of these data the same SPSS program called linear regression and command again: **Statistics; Regression; Linear**. The software **SPSS 8 for windows 99** provides us with the following three subtables: **(1) Model Summary, (2) ANOVA, (3) coefficients** (Table 4).

Table 4.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848[a]	.719	.701	5.4498

a. Predictors: (Constant), VAR00003, VAR00002

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2429.764	2	1214.882	40.905	.000[a]
	Residual	950.407	32	29.700		
	Total	3380.171	34			

a. Predictors: (Constant), VAR00003, VAR00002

b. Dependent Variable: VAR00001

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.547	4.233		-.366	.717
	VAR00002	1.701	.269	.653	6.312	.000
	VAR00003	.426	.134	.330	3.185	.003

a. Dependent Variable: VAR00001

(1)Model Summary shows r, here called the multiple r, The corresponding "multiple r square", otherwise called coefficient of determination, of 0.719 indicates that 71.9 % of the variation in the y variable is explained by variation in the two x variables. Interestingly, the multiple r square is a bit larger than the simple r square (0.719 and 0.618). Information is thus given about the perfection of the model. After the first step 61.8 % of variation is explained by the regression model, after the second no less than 71.9% is explained by it. The addition of age to the model produces 71.9-63= 8.9% extra explanation of the variance in the y variable, the effect of the new laxative. The interpretation of the r^2 – value is similar to that in simple linear regression. If $r^2 = 0$, then no correlation exists, the x-values determines the y-values no way. If $r^2 = 1$, then the correlation is 100%, we are absolutely sure about the y-value if we know the x-values. If $r^2 = 0.5$, the 50% correlation exists. In our case $r^2 = 0.719 = 72\%$. The x-values determine the y-values by 72% certainty. We have 28% uncertainty =noise = (SE of $r^2 = 1-r^2$).

Before going further we have to consider the hazard of collinearity, which is the situation where two x variables are highly correlated. One naive though common way in which collinearity is introduced into the data, is through inclusion of x variables that are actually the same measures under different names. This is, obviously, not so with bisacodyl effect and age. Nonetheless, we measure the presence of collinearity by calculating the simple correlation coefficient between the x variables before doing anything more. In our case r between x_1 variables and x_2 variables is 0.425, and so we don´t have to worry about (multi)collinearity (r>0.90).

(2)ANOVA is used to test whether r is significantly larger than 0.00. Again SS regression (by regression explained variance) is divided by SS residual (unexplained variance), the total variance being SS regression + SS residual. The division sum "304.570 / SS total" yields 0.719 = r square, Called R square by SPSS. If r^2 is significantly different from the 0, then a regression plane like the one from Figure 5 is no accident. If r^2 is significantly larger than 0 like here, then the data are closer to the regression plane than could happen by accident.

(3)Coefficients again shows the real regression equation. The intercept a is given by the (constant). The b values are the unstandardized regression coefficients of the x_1 and x_2 variables.

The regression equation is thus as follows

$y = - 1.547 + 1.701 \,.\, x_1 + 0.426 \,.\, x_2$

new laxative = - 1.547 + 1.701 . bisacodyl + 0.426 . age

In addition to unstandardized coefficients, standardized coefficients are given. For that purpose SS y is taken to be 1. Then r = b. Instead of testing the null hypothesis that r= 0 we can now test that various $b_i = 0$, and use for that purpose t-test. As both bisacodyl and age are significantly correlated with the y variable (the efficacy of the new laxative), both x variables are independent predictors of the efficacy of the new laxative.

6. PURPOSES OF LINEAR REGRESSION ANALYSIS

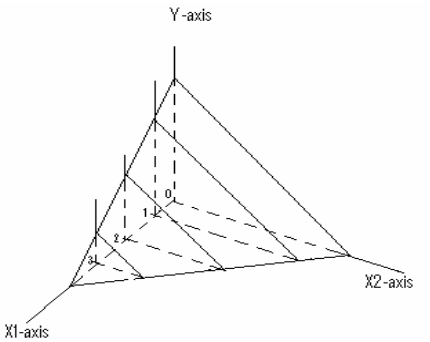


Figure 6. Regression plane.

In summary, multiple regression-analysis with 3 variables and the equation formula $y = a + b_1 x_1 + b_2 x_2$, can be illustrated by a regression plane, the best fit plane for the scattered data (Figure 6). A p-value <0.0001 means that the data are a lot closer to the regression plane than could happen by accident. If more than 3 variables are in the model, then the model becomes multidimensional, and a graph is impossible, but the principle remains the same.

Multiple linear regression analysis is used for different purposes (see also the next chapter). The above example of two x-variables is an example where multiple linear regression is used in a controlled clinical crossover trial in order to provide more *precision* in the data. With a single x-variable the R^2 -value = 63%, with two x-variables the R^2 -value = 72%. Obviously, the level of certainty for making prediction about the y-variable increases by 72% - 63% = 9%, if a second x-variable is added to the data. Chapter 17 will give additional examples of this purpose. Another common purpose for its use is *exploratory* purposes. We search for significant predictors = independent determinants of the y-variable, and include multiple x-variables in the model. Subsequently, we asses which of the x-variables included are the statistically significant predictors of the y-variable according to the model

$$y = a+b_1 x_1 + b_2 x_2 +\ldots\ldots\ldots b_{10} x_{10}$$

The b-values are the partial correlation coefficients, and are used to test the strength of the correlation. If b_1 t/m b_{10} are significantly < / > 0, then the corresponding x-variable is a significant predictor of the y-variable. The different x-variables can be added to the model one by one (stepwise, step-up), or all together. If added all together, we remove the insignificant ones starting with the one with the largest p-value (stepwise, step down). In practice the step-up and step-down method will produce rather similar results. If none of the x-variables produces a significant b-value, but the overall R^2 –value is significantly different from 0, we have to conclude that none of the x-variables is an independent determinant of the y-variable, yet the y-value is significantly dependent on all of the x-variables.

Two more purposes of linear regression are the assessment of *confounding* and *interaction*. These purposes will be discussed as an introduction in the next chapter, and more fully in the Chapters 19 and 20.

7. ANOTHER REAL DATA EXAMPLE OF MULTIPLE LINEAR REGRESSION (EXPLORATORY PURPOSE)

We want to study "Independent determinants of quality of life of patients with angina pectoris". Note this is an observational rather than interventional study. We give the example because these kinds of data are often obtained as secondary data from interventional studies.

y-variable= index of quality of life of patients with stable angina pectoris
x-variables=1.Age
2.Gender
3.Rhythm disturbances
4.Peripheral vascular disease
5.Concomitant calcium channel blockers
6.Concomitant beta blockers
7.NYHA-classification
8.Smoking
9.body mass index
10.hypercholesterolemia
11.hypertension
12.diabetes mellitus

Index of quality of life = $a + b_1$ (age) + b_2 (gender) + …… b_{12} (diabetes)

Correlation between independent variables may be correlated but not too closely: e.g. body mass index, body weight, body length should not be included all three. We used single linear regression for assessing this correlation, otherwise called multicollinearity (Table 5).

Table 5. correlation matrix in order to test multicollinearity in the regression analysis, P-values are given

	age /	gender /	rhythm /	vasc dis /	ccb /	bb /	NYHA /	smoking /	bmi /	chol /	hypt
gender	0.19	1.00									
rhythm	0.12	ns	1.00								
vasc dis	0.14	ns	ns	1.00							
ccb	0.24	ns	0.07	ns	1.00						
bb	0.33	ns	ns	ns	0.07	1.00					
NYHA	0.22	ns	ns	0.07	0.07	ns	1.00				
smoking	–0.12	ns	0.09	0.07	0.08	ns	0.50	1.00			
bmi	0.13	ns	ns	ns	ns	0.10	-0.07	0.62	1.00		
chol	0.15	ns	ns	0.12	0.09	ns	0.08	0.09	ns	1.00	
hypt	0.09	ns	0.08	ns	0.10	0.09	0.09	0.09	0.07	0.41	1.00
diabetes	0.12	ns	0.09	0.10	ns	0.08	ns	0.11	0.12	0.10	0.11

vasc dis= peripheral vascular disease; ccb= calcium channel blocker therapy; bb= beta-blocker therapy; bmi= body mass index; hypt= hypertension; ns= not statistically significantly correlated (Pearson's correlation p-value>0.05).

Table 6 shows the b-values that are not significantly different from 0. They are removed from the model. This procedure is called the step-down method (the step-up method includes the variables one by one, while removing those with an insignificant b-value). Table 7 summarizes the significant b-values. Conclusions: The higher the NYHA class the lower quality of life (Figures 7 and 8). Smokers, obese subjects, and patients with concomitant hypertension have lower quality of life. Patients with hypercholesterolemia or diabetes mellitus have better quality of life. The latter two categories may have early endothelial dysfunction and may have significant angina pectoris with fairly intact coronary arteries. An alternative interpretation is that they have better quality of life because they better enjoy life despite a not so healthy lifestyle. This uncertainty about the cause of relationship established illustrates uncertainties produced by regression analyses. Regression analyses often establish relationships that are not causal, but rather induced by some unknown common factor.

Table 6. B-values used to test correlation, step down method

x-variable	regression coëfficient (B)	standard error	test (T)	Significance level (p-value)

Age	-0.03	0.04	0.8	0.39
Gender	0.01	0.05	0.5	0.72
Rhythm disturbances	-0.04	0.04	1.0	0.28
Peripheral vascular disease	-0.00	0.01	0.1	0.97
Calcium channel blockers	0.00	0.01	0.1	0.99
beta blockers	0.03	0.04	0.7	0.43

Table 7. B-values to test correlation, step down method

x -variable	regression coëfficient (B)	standard error	test stat (T)	Significance level (p-value)
NYHA-classification	-0.08	0.03	2.3	0.02
Smoking	-0.06	0.04	1.6	0.08
body mass index	-0.07	0.03	2.1	0.04
hypercholesterolemia	0.07	0.03	2.2	0.03
hypertension	-0.08	0.03	2.3	0.02
diabetes mellitus	0.06	0.03	2.0	0.05

NYHA = New York Heart Association.

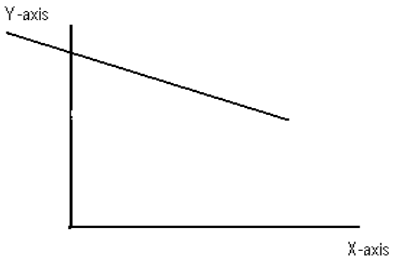


Figure 7. A negative b-value indicates: if x >, then y <.

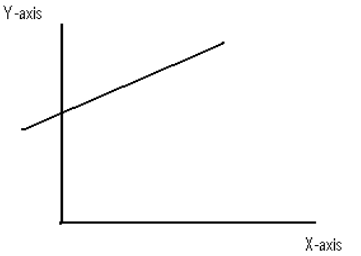


Figure 8. A positive b-value indicates: if x > then y >.

8. IT MAY BE HARD TO DEFINE WHAT IS DETERMINED BY WHAT, MULTIPLE AND MULTIVARIATE REGRESSION

It may be sometimes hard in a linear regression to define what is determined by what, or, in other words, what are the dependent (y-values) and the independent variables (x-values). Generally, it is helpful to consider as independent determinants "causal – factors" determining the result, while the result is the dependent variable, otherwise called outcome variable. Independent variables are currently often called exposure variables or indicator variables. In regression

analyses of clinical trials the treatments modalities, in addition to patients' characteristics, are often independent variables. As examples we give two patient series with multiple variables:

(1) Type operation
(2) Type surgeon
(3) Complications yes/no
(4) Gender patients
(5) Age patients
(6) Required time for recovery,

(2) may determine (1), (3), and (6), but not (4) and (5),
(4) and (5) maybe (1),
(1) does not determine (4) and (5).

In another patient series the variables are:

(1) 2 types of anesthesia
(2) Pain scores
(3) Complications yes/no
(4) Gender patients
(5) Age patients
(6) Comorbidity preoperatively
(7) Quality of life after surgery

(1) determines (2) and maybe (3) and (7), but not (4), (5), and (6),
(4), (5), and (6) may determine (1) and maybe also (2), (3), and (7).
Regression can be nonsense and still produce significant results, e.g., if you let (1) determine (4), (5), and (6).

Mostly, a single y-variable and multiple x-variables are included in a regression analysis, and this is what we call a multiple regression analysis. In the reports the term multivariate analysis is often erroneously used for these models. The term multivariate analysis refers to models that include more than a single y-variable and the analysis is then called multivariate analysis of variance (MANOVA). a correct alternative term for multiple regression analysis is, thus, univariate analyses with multiple x-variables (independent variables).

9. LIMITATIONS OF LINEAR REGRESSION

The limitations of multiple regressions are reviewed in the above text, but a summary is given:

1. The risk of multicollinearity.
2. The requirement of homoscedasticity.
3. The spread around the y-values is inthe form of equal Gaussian curves, if not then Rank correlation according to Spearman should be performed.
4. A linear correlation between x and y exists.

5. The risk of confounding.
6. The risk of interaction.

10. CONCLUSIONS

If the above information is too much, don't be disappointed: multiple linear regression analysis and its extensions like logistic regression and Cox's proportional hazard model are not as important for clinical trials as it is for observational research:

1. Regression analysis assesses associations not causalities.
2. Clinical trials assess causal relationships.
3. We believe in causality if factor is introduced and gives rise to a particular outcome.
4. Always air of uncertainty with regression analysis

Multiple linear regression is interesting, but, in the context of clinical trials mostly just exploratory.

CHAPTER 14

SUBGROUP ANALYSIS USING MULTIPLE LINEAR REGRESSION: CONFOUNDING, INTERACTION, SYNERGISM

1. INTRODUCTION

When the size of the study permits, important demographic or baseline value-defined subgroups of patients can be studied for unusually large or small efficacy responses; e.g. comparison of effects by age, sex; by severity or prognostic groups. Naturally, such analyses are not intended to "salvage" an otherwise negative study, but may be may be helpful in refining patient or dose selection for subsequent studies.[1]

Most studies have insufficient size to assess efficacy meaningfully in subgroups of patients. Instead a regression model for the primary or secondary efficacy-variables can be used to evaluate whether specific variables are confounders for the treatment effect, and whether the treatment effect interacts with specific covariates. The particular (statistical) regression model chosen, depends on the nature of the efficacy variables, and the covariates to be considered should be meaningful according to the current state of knowledge. In particular, when studying interactions, the results of the regression analysis are more valid when complemented by additional exploratory analyses within relevant subgroups of patients or within strata defined by the covariates.

In this chapter we will discuss the multiple linear regression model which is appropriate, for continuous efficacy variables, such as blood pressures or lipid levels (as discussed in chapter 2). Regression models for dichotomous efficacy variables (logistic regression[2]), and for survival data (Cox regression[3]) will not be assessed here. However, the principles underlying all of these models are to some extent equivalent.

2. EXAMPLE

As an example of the use of a regression model we consider trials such as those conducted to evaluate the efficacy of statins (HMG-CoA reductase inhibitors) to lower lipid levels in patients with atherosclerosis.[4] In unselected populations statins were extremely effective in lowering LDL cholesterol (LDL), but the question whether the efficacy depended on baseline LDL level was unanswered. Of course this could be answered by comparing efficacy in selected subgroups of patients with baseline *low*, *intermediate*, and *high* LDL levels, but a regression model could be used as well, and sometimes provides better sensitivity.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 175–184.

Consider a randomized clinical trial such as Regress.[4] In this trial 884 patients with documented coronary atherosclerosis and total cholesterol between 4 and 8 mmol/L were randomized to either two-year pravastatin or placebo treatment. Efficacy of treatment was assessed by the fall in LDL cholesterol after two year treatment. In the n_1=438 patients who received pravastatin mean LDL cholesterol fell by $\bar{x}_1 = 1.2324$ mmol/L (standard deviation, $S_1 = 0.68$). In the $n_0 = 422$ available patients who received placebo, the mean LDL cholesterol fell by $\bar{x}_0 = -0.0376$ mmol/L ($S_0 = 0.589$). Consequently, the efficacy of pravastatin was 1.2324 - -0.0376 = 1.2700 mmol/L LDL-decrease in two years with standard error (SE) 0.043 mmol/l, and the 95% confidence interval (ci) of the efficacy quantification ran from 1.185 to 1.355.
In a random patient with coronary atherosclerosis and total cholesterol in between 4 and 8 mmol/L, pravastatin produces a better reduction in LDL cholesterol than does placebo by 1.27 mmol/L. However, a patient with 8 mmol/L total cholesterol level may better benefit than a patient with 4 mmol/L at baseline may do. Multiple linear regression can be applied to assess this question.

3. MODEL (FIGURE 1)

We first introduce some notation: the dependent variable Y_i is the amount of LDL decrease observed in patient i (i=1,...,884), and the independent variable or covariate X_{1i} is an indicator variable, indicating whether patient i received pravastatin ($X_{1i} = 1$) or not ($X_{1i} = 0$). We define the linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + e_i, \quad (1)$$

where β_0 is the intercept, and β_1 the slope of the regression line and e_i is a residual variation term, which is assumed to be normally distributed with variance σ_e^2.

When X_{1i} is either zero or one, the usual estimates b_0, b_1, and S_e^2 of β_0, β_1, and $=_e^2$ are:

$$b_0 = \bar{x}_0 = -0.0376, \quad b_1 = \bar{x}_1 - \bar{x}_0 = 1.2700, \text{ and}$$

$$S_e^2 = \frac{(n_1 - 1) S_1^2 + (n_0 - 1) S_0^2}{n_1 + n_0 - 2} = 0.4058,$$

which are exactly the same statistics as used in the t-test procedure one would normally employ in this situation. The quantification of the efficacy is thus given by b_1 and it has the same value and the same standard error and confidence interval as above. In Figure 1 the linear regression line is illustrated.
Note: b and s are the best estimates, otherwise called best fits, of β and σ.

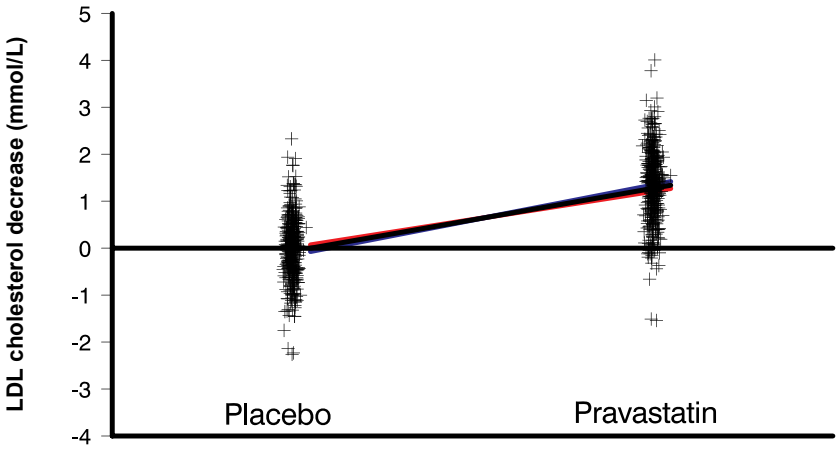


Figure 1. The linear regression line is illustrated.

By using this regression model the following assumptions are made.

1. The relation between Y and X is linear. When X can attain only two values, this assumption is naturally valid, but, otherwise, this is not necessarily so.
2. The distribution of the residual term e_i is normal with mean zero and variance σ^2_e.
3. The variance of the distribution of e, σ^2_e, is the same for $X_1 = 0$ and for $X_1 = 1$: homoscedasticity.
4. The residual term e_i is independent of X_{1i}.

The object of regression modeling in clinical trials is to evaluate whether the efficacy quantification b_1 (I.) can be made more precise by taking covariates into consideration, (II.) is confounded by covariates, and (III.) interacts with covariates (synergism).
Increased precision (I.) is attained, and confounding (II.) can be studied by extending the regression model with a second independent variable X_2:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, (2).$$

This multiple regression model has the same underlying assumptions as the above linear regression model (1) except for the assumption that e_i is independent not only of X_1 but also of X_2. There is no need to assume that X_1 and X_2 are strictly independent, but the association must not be too strong (multicollinearity).

4. (I.) INCREASED PRECISION OF EFFICACY (FIGURE 2)

When X_2 is independent of X_1 and is associated with Y (thus $b_2 \neq 0$), the estimate b_1 of the model in equation (2) will be the same as the estimate b_1 of the model in equation (1), but its precision will be increased, as indicated by a smaller standard error.

This is a common case in randomized clinical trials. The randomization will ensure that no inbalances exist between the two treatment groups with respect to covariates such as X_2, and consequently X_2 will be independent of the treatment variable X_1. There are often many candidates for inclusion as covariates in the multiple regression model, but the choice should be made a priori and specified in the protocol. When the dependent variable is a change score, as in our example, the baseline level is the first candidate to consider because it is almost surely associated with the change score Y. Figure 2 shows the relationship between result of treatment and baseline values as demonstrated by scatterplots and linear regression lines for each treatment separately. The multiple linear regression model in equation (2) is appropriate for testing the contribution of baseline variability to the overall variability in the data.

Since X_2 is independent of X_1, inclusion of X_2 in the model must lead to a decreased variance S_e^2: some differences between patients with respect to the LDL decrease, are attributed to baseline LDL levels. Thus there will be less residual variation. Since the standard error of b_1 is a monotonic positive function of S_e^2, a decrease of S_e^2 leads to a smaller standard error of b_1. Thus by including baseline LDL levels in the regression model, the efficacy of pravastatin lowering is estimated more precisely. This rule, however, only applies to large data-sets. With every additional covariate in the model an extra regression weight must be estimated, and since S_e^2 is an inverse function of the number of covariates in the model, too many covariates in the model will lead to decreased precision.

In our example the mean baseline LDL levels (X_2) were 4.32 (SD 0.78) and 4.29 (SD 0.78) in the placebo and pravastatin treatment groups (X_1) (p=0.60); hence X_1 and X_2 were independent. The baseline LDL levels were, however, associated with the LDL-changes (Y): b_2=0.41 (SE 0.024), p<0.0001. Consequently, the estimated efficacy was (almost) the same as before, but it had a somewhat smaller standard error, and is, thus, more precise:

with baseline LDL cholesterol levels:	b_1 = 1.27 (SE 0.037)
without baseline LDL cholesterol levels:	b_1 = 1.27 (SE 0.043)

Additional examples of regression modelling for improved precision are given in chapter 15.

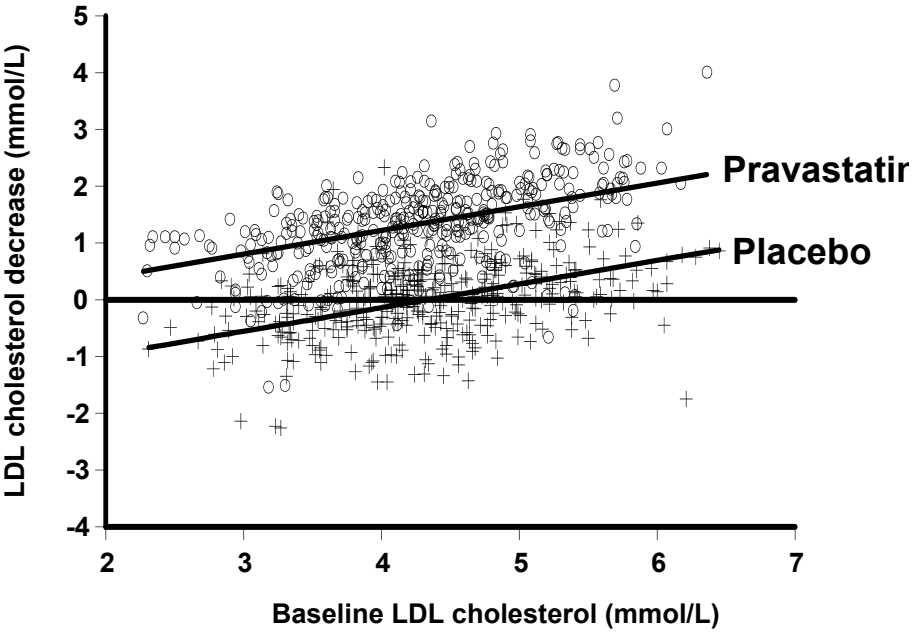


Figure 2. Scatterplots and linear regression lines of baseline LDL cholesterol and LDL cholesterol decrease after treatment, separately for placebo and for pravastatin treatments.

Note: in contrast to the linear regression models the efficacy estimates of non-linear regression models (e.g. logistic[2] and Cox regression[3]) do not remain the same in this case. When using logistic or Cox regression it is, therefore, imperative to report the log odds ratio or log hazard ratio of treatments compared, together with the covariates in the model.

5. (II.) CONFOUNDING

In randomized clinical trials confounding plays a minor role in the data. The randomization will ensure that no covariate of the efficacy variable will also be associated with the randomized treatment. If, however, the randomization fails for a particular variable, which is already known to be an important covariate of the efficacy variable, such a variable is a confounder and adjustment of the efficacy estimate should be attempted. This is done by using the same (linear) regression model as given in equation (2). The adjusted efficacy estimate may become smaller or larger than the unadjusted estimate, depending on the direction of the associations of the confounder with the randomized treatment and the efficacy variable. Let b_1 and b_1^* denote the unadjusted and the adjusted efficacy estimate, and let r_{xz} and r_{yz} be the

correlations of the confounder (z) with the randomized treatment (x) and the efficacy variable (y), then the following will hold:

if	$r_{xz}>0$ and $r_{yz}>0$	then	$\lvert b_1^* \rvert < \lvert b_1 \rvert$,
if	$r_{xz}>0$ and $r_{yz}<0$	then	$\lvert b_1^* \rvert > \lvert b_1 \rvert$,
if	$r_{xz}<0$ and $r_{yz}<0$	then	$\lvert b_1^* \rvert < \lvert b_1 \rvert$,
if	$r_{xz}<0$ and $r_{yz}>0$	then	$\lvert b_1^* \rvert > \lvert b_1 \rvert$,

Notice the possibility that the unadjusted efficacy estimate b_1 is zero whereas the adjusted estimate b_1^* is unequal to zero: an efficacy-difference between treatments may be masked by confounding.
In clinical trials it is sensible to check the balance between treatment groups of all known covariates of the efficacy variable. In most trials there are many more covariates and one should be careful to consider as a confounder a covariate which was not reported in the literature before.

6. (III.) INTERACTION AND SYNERGISM

A special kind of covariate is the interaction of the randomized treatment with some other covariate. This interaction is, by definition, associated with the randomized treatment, and possibly with the efficacy variable if the efficacy differs between treatments. In contrast to the discussion above, the focus of the statistical analysis is not on the change of b_1 by including an interaction in the model, but the regression weight of the interaction variable itself. When this regression weight is unequal to zero, this points to the existence of patient-subgroups for which the efficacy of treatment differs significantly.
An example is again provided by the Regress trial.[4] The primary effect variable was the decrease of the average diameter of the coronary arteries after two years of treatment. The average decrease was 0.057 mm (standard deviation (SD) 0.194) in the pravastatin group, and it was 0.117 mm (SD 0.212) in the placebo group (t-test: significance of difference at $p<0.001$); thus the efficacy estimate b_1 was 0.060 (standard error SE = 0.016). Calcium channel blockers (CCB) were given to 60% of the placebo patients, and 59% of the pravastatin patients (chi-square: p=0.84): thus CCB treatment was not a confounder variable. Also, CCB medication was not associated with diameter decrease (p = 0.62). In the patients who did not receive concomitant CCB medication, the diameter decreases were 0.097 (SD 0.20) and 0.088 (SD 0.19) in patients receiving placebo and pravastatin, respectively (p = 0.71). In patients who did receive CCB medication, the diameter decreases were 0.130 (SD 0.22) and 0.035 (SD 0.19), respectively ($p<0.001$). Thus, pravastatin-efficacy was, on average, 0.097-0.088 = 0.009 mm in patients without CCB medication, and 0.130-0.035 = 0.095 in patients with CCB medication.

This difference was statistically significant (interaction test: p=0.011). We used the following linear regression model for this test. Let X_{1i}=1 denote that patient i received pravastatin (X_{1i}=0, if not), let X_{2i}=1 denote that patient i received CCB medication (X_{2i} = 0, if not), and let X_{3i} = X_{1i} x X_{2i}:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i .$$

The estimates were: b_3 = 0.085 (SE 0.033), b_2 = -0.033 (SE 0.023), and b_1 = 0.009 (SE 0.026). Notice that b_1 changed dramatically by including the interaction term X_3 in the linear model; this is a general feature of regression models with interaction terms: the corresponding main-effects (b_1 and b_2) cannot be interpreted independently of the interaction term. Another consequence is that the efficacy estimate no longer exists, but several estimates do exist: in our case there are different efficacy-estimates for patients with (b_1+b_3 = 0.009+0.085 = 0.094) and without CCB medication (b_1 = 0.009). In the practice of clinical trials interactions are usually investigated in an exploratory fashion. When interaction is demonstrated in this way, its existence should be confirmed in a novel prospective clinical trial. Additional examples of regression modeling for interaction effects are given in chapter 17.

7. ESTIMATION, AND HYPOTHESIS TESTING

Standard statistical computer programs like SPSS and SAS (and many others) contain modules that perform regression analysis for linear and many non-linear models such as logistic and Cox regression. The standard method to estimate the linear regression weights (and the residual standard deviation σ_e) is to minimize the squared distances between the data and the estimated regression line: the least squares method. For non-linear models, the maximum likelihood method is employed, but these are equivalent methods. The output of these estimation methods are the estimated regression weights (and the residual standard deviation σ_e) and their standard errors. It is important that the correlations between the covariates in the model are not too large (i.e. multicollinearity), but if these are too large, this will become clear by absurd regression weights, and very large standard errors. If this occurs, one or more covariates must be removed from the model.

Under the null hypothesis that β equals zero, the ratio of the estimated regression weight b and its standard error is distributed as a student's t statistic in the linear model, and this can be used to derive the p-value or the 95% confidence interval in the usual way. For non-linear models, the squared ratio of b and its standard error is called the Wald statistic which is chi-squared distributed. Alternatives for the Wald statistic are the score and likelihood ratio statistics[5], but these give the same results except in highly unusual circumstances; if they differ, the score and likelihood statistics are better than the Wald statistic.

The power of these statistical tests is a sensitive function of the number of patients in the trial. Naturally, there is less opportunity for modeling in a small trial than in a

large trial. There is no general rule about which sample sizes are required for sensible regression modeling, but one rule-of-thumb is that at least ten times as many patients are required as the number of covariates in the model.

8. GOODNESS-OF-FIT

For the linear model the central assumptions are (1.) the assumed linearity of the relation between Y and X, and (2.) the normal distribution of the residual term e independent of all covariates and with homogeneous variance. The first step in checking these assumptions is by looking at the data. The linearity of the relation between Y and X, for instance, can be inspected by looking at the scatter-plot between Y and X. A nonlinear relation between Y and X will show itself as systematic deviation from a straight line. When the relation is nonlinear, either Y or X or both may be transformed appropriately; most often used are the logarithmic transformation $X^*=\ln(X)$ and the power transformation $X^*=X^p$ (e.g. the squared root transformation where p = 0.5). At this stage subjective judgments necessarily enter the statistical analysis, because the decision about the appropriate transformation is not well founded on statistical arguments. A few tools that may help, are the following.

1. The optimal power-transformation (X^p) may be estimated using the Box-Cox algorithm.[3] This may yield, however, difficult and unpractical power-transforms.
2. A 'better' model produces better correlations. When one compares two different models, the better of the two leads to a smallest residual variance (S^2_e) or highest multiple correlation coefficient (R): $S^2_e=[(n-1)/(n-k)](1-R^2)S^2_y$, where k is the number of covariates in the model.
3. Choosing an appropriate transformation may be enhanced by modelling the relation between Y and X as a polynomial function of X: $Y=b_0+b_1X+b_2X^2+b_3X^3+....$. When the relation is strictly linear then $b_2 = b_3 =...= 0$, and this can be tested statistically in the usual way. Obviously, the order of the polynomial function is unknown, but one rarely needs to investigate fourth or higher orders.
4. Finally, there exists the possibility to model the association between Y and X nonparametrically using various modern smoothing techniques.

The assumed normal distribution of the residual term can be checked by inspecting the histogram of e. The estimation method and the hypothesis testing are quite robust against skewed distributions of the residual term, but it is sensible to check for extreme skewness and the occurrence of important outlying data-points. Visual inspection is usually sufficient but one may check the distribution statistically with the Kolmogorov-Smirnov test (see also chapter 25).

More important is the assumption of homogeneity of the residual variance S^2: this entails that the variation of e is more or less the same for all values of X. One may check this visually by inspecting the scatterplot of e (or Y) versus X. If heterogeneity is present, again an appropriate transformation of Y may help. If the ratio of S_e / y is equal for various levels of X, the logarithmic transformation Y* = ln(Y) may help, and if S_e^2 / y^2 is equal for various levels of X, the square-root transformation is appropriate: $Y^* = (Y)^{0.5}$. The independence of the residual term e of all covariates X in the model can be tested with the Durbin-Watson test.
In the logistic regression model the most important underlying assumption is the assumed logistic form of the function linking the covariates to the binary efficacy variable. When not all relevant covariates are in the model, it can be shown that the link-function is not logistic. One way to statistically test this, is by using the Hosmer-Lemeshow test[2]. But if the logistic regression model does not fit, this is of little consequence because this usually points to missing covariates, and these are often not available. In Cox regression, the cardinal underlying assumption is the assumed proportionality of the hazard rates. There are several statistical tests for this assumption; if proportionality does not hold, accelerated failure time models can be used, or the time axis may be partitioned into several periods in which proportionality does hold.

9. SELECTION PROCEDURES

In clinical trials usually many variables are sampled, and often many of these are candidates for inclusion in the regression model. A major problem is the selection of a subset of variables to include in the regression model. By far preferable is to select a (small) set of candidate variables on clinical and theoretical grounds, but if that is not possible a few rules are helpful in the selection process.

1. If the number of covariates is not too large, it is best not to use any selection at all, but simply include all candidates in the regression model. Often it is necessary to shrink the regression weights, using, for instance, a penalty function.
2. If the number of covariates is very large, backward selection methods are preferable to forward selection models. This is usually done according to the p-value or the size of the test-statistic-value.
3. Since the overriding interest of the regression modelling is the estimation of the efficacy of the randomized treatments, the safest course is to be liberal about including covariates in the model: use a p-value of 0.10 or even 0.20 to include covariates in the model.

10. MAIN CONCLUSION

The regular statistical analysis of the data of clinical trials should be extended by (exploratory) analysis if the existence of subgroups of patients for which the efficacy estimate differs, is suspected. An efficient way of doing this is by the use of

regression analysis. If such subgroups are identified, the exploratory nature of the regression analysis should be emphasized and the subgroup issue should be further assessed in subsequent independent and prospective data-sets.

11. REFERENCES

1. Department of Health and Human Services, Food and Drug Administration. International Conference on Harmonisation; Guidance on Statistical Principles for Clinical Trials Availability. Federal Register, 63 (179), 1998: 49583-49598.
2. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley, 1989.
3. Box Cox, Statistical Software, University Leyden, Netherlands, 1999.
4. Jukema AJ, Zwinderman AH, et al for the REGRESS study group. Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elevated serum cholesterol levels. The Regression Growth Evaluation Statin Study (REGRESS). Circulation. 1995; 91: 2528-40.
5. Rao CR. Linear Statistical Inference and Its Applications. New York: Wiley, 1973.

CHAPTER 15

CURVILINEAR REGRESSION

1. INTRODUCTION

Polynomial analysis is an extension of simple linear regression, where a model is used to allow for the existence of a systematic dependence of the dependent y variable (blood pressure) on the independent x variable (time) different from a linear dependence. Polynomial extension from the basic model can be done as follows:

$y = a + bx$	(first order) linear relationship
$y = a + bx + cx^2$	(second order) parabolic relationship
$y = a + bx + cx^2 + dx^3$	(third order) hyperbolic relationship
$y = a + bx + cx^2 + dx^3 + ex^4$	(fourth order) sinusoidal relationship

where a is the intercept and b, c, d, and e are the partial regression coefficients. Statistical software can be used to calculate for the data the regression line that provides the best fit for the data. In addition, regression lines of higher than 4 orders can be calculated. Fourier analysis is a more traditional way of analyzing these type of data, and is given by the function

$f(x) = p + q_1 \cos(x) + ..+q_n \cos n(x) + r_1 \sin(x) +..+ r_n \sin n(x)$
with p, $q_1...q_n$, and $r_1...r_n$ = constants for the best fit of the given data.

As an example, ambulatory blood pressure monitoring (ABPM) using light weight automated portable equipment is given. ABPM has greatly contributed to our understanding of the circadian patterns of blood pressures in individual patients[1] as well as to the study of effects of antihypertensive drugs in groups of patients.[2] However, a problem is that ABPM data using mean values of arbitrarily separated daytime hours are poorly reproducible[3,4], undermining the validity of this diagnostic tool. Previous studies have demonstrated that both in normo-[5] and in hypertensive groups[6] time is a more powerful source of variation in 24 hour ABPM data than were other sources of variation (between $P<0.01$ and <0.001 versus between not significant and <0.01). This reflects the importance of the circadian rhythm in the interpretation of ABPM data, and the need for an assessment that accounts for this very rhythm more adequately than does the means of separated daytime hours. We also demonstrated that polynomial curves can be produced of ABPM data from both normo-[5] and hypertensive[6] groups, and that these polynomial curves are within the 95% confidence intervals of the sample means. However, intra-individual reproducibility of this approach has not been assessed, and is a prerequisite for further implementing this approach.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 185–197.

In this chapter we describe polynomial analysis of ABPM data, and test the hypothesis that it is better reproducible and that this is so, not only with means of populations, but also with individual data. For the estimate of reproducibility duplicate standard deviations as well as intra-class correlations are calculated of ABPM data from untreated mildly hypertensive patients who underwent ABPM for 24 hours twice, 1 week interval.

2. METHODS, STATISTICAL MODEL

Ten patients, 6 females and 4 males, who had given their informed consent, participated in the study. Each patient had been examined at our outpatient clinic. Age varied from 33 to 52 years of age (mean 42 years), body mass index from 20 to 31 kg/m (mean 29 kg/m). Patients were either housewife or actively employed throughout the study and had no other diseases. Previously treated patients had a washout period of at least 8 weeks before they were included in the study. All patients were included if untreated diastolic blood pressure was repeatedly between 90 and 100 mm Hg and systolic blood pressure less than 170 mm Hg.
In all of the patients ABPM consisted of measurements every 60 minutes for 24 hours with a validated[7] light weight automated portable equipment (Space Lab Medical Inc, Redmond WA, model 90207). In the meantime patients performed their usual daily activities.
We define the dependent variable, the blood pressure recording at hour t, and, subsequently, model it as a function of hour t, hour t squared, hour t to the power 3, hour t to the power 4, and so on. The a- and b-values are constants for the best fit of the given data, and are also called the regression weights.

$$\text{Blood pressure at hour } t = a + b_1 (\text{hour } t) + b_2 (\text{hour } t)^2 + b_3 (\text{hour } t)^3 + b_4 (\text{hour } t)^4 + \ldots.$$

If we use Fourier analysis instead the equation is

$$\text{Blood pressure at hour } t = p + q_1 \cos (\text{hour } t) + .. + q_n \cos n(\text{hour } t) + r_1 \sin (\text{hour } t) + .. + r_n \sin n(\text{hour } t)$$

with p, $q_{1\text{-}n}$ and $r_{1\text{-}n}$ being constants for the best fit of the given data.

Reproducibility of ABPM was studied in the ten patients by performing 24 hour ABPM in each of them twice, intervals at least 1 week. Reproducibility of the duplicate data, as obtained, were assessed both by quantifying reproducibility of means of the population, and of the individual data.

REPRODUCIBILITY OF MEANS OF THE POPULATION

For this purpose we used duplicate standards deviation (Duplicate SD) and intra-class correlation (ρ_I).[8]

Duplicate SD was calculated according to Duplicate $SD = \sqrt{\frac{\sum (x_1 - x_2)^2}{2n}}$, where x_1 and x_2 are individual data during 1st and 2nd tests, and n= 240 (10 times 24 duplicate observations).

Intra-class correlation (ρ_I) is another approach for the estimate of replicability of repeated measures in one subject, and is calculated according to

$$\rho_1 = \frac{\sigma^2 \bar{x}_1 - \sigma^2 \bar{x}_2 / \bar{x}_1}{\sigma^2 \bar{x}_1}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the 240 values during test 1 and test 2 respectively, and $\sigma^2 \bar{x}_2 / \bar{x}_1$ is the variance of $\bar{x}_2$ given $\bar{x}_1$, and

$$\sigma^2 \frac{\bar{x}_2}{\bar{x}_1} = \sigma^2 \bar{x}_1 - \frac{(\bar{x}_1 - \bar{x}_2)^2}{4}.$$

A slightly different method to calculate intraclass correlations is described in chapter 26.

Note: Greek symbols like σ instead of s and ρ instead of r are often used in statistics. They are used to indicate population parameters instead of sample parameters.

REPRODUCIBILITY OF INDIVIDUAL DATA

For this purpose we similarly used duplicate standards deviation (SD) and intra-class correlation (ρ).

Duplicate SD was calculated according $SD = \sqrt{\frac{\sum (x_1 - x_2)^2}{2n}}$ where x_1 and x_2 are individual data during 1st and 2nd tests, and n= 24 (24 duplicate observations per patient).

Intra-class correlation (ρ_I) was calculated according to

$$\rho_1 = \frac{\sigma^2 \bar{x}_1 - \sigma^2 \bar{x}_2 / \bar{x}_1}{\sigma^2 \bar{x}_1}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the 24 values during test 1 and test 2 respectively, and $\sigma^2 \bar{x}_2 / \bar{x}_1$ is the variance of $\bar{x}_2$ given $\bar{x}_1$, and

$$\sigma^2 \frac{\bar{x}_2}{\bar{x}_1} = \sigma^2 \bar{x}_1 - \frac{(\bar{x}_1 - \bar{x}_2)^2}{4}$$

Calculations were performed using SPSS statistical software, polynomial curves were drawn using Harvard Graphics 3.[9,10] Under the assumption of standard deviations of 25 % and intraclass correlations of + 0.7, at least 240 duplicate observations had to be included to obtain a regression analysis with a statistical power of 80% and a 5 % significance level. And so, it seemed appropriate to include hourly data of at least 10 patients tested twice for 24 hours. Paired means,

Duplicate SDs and intraclass correlations were statistically tested by t-tests, F tests, or McNemar's chi-square tests, whenever appropriate.

3. RESULTS

REPRODUCIBILITY OF MEANS OF POPULATION

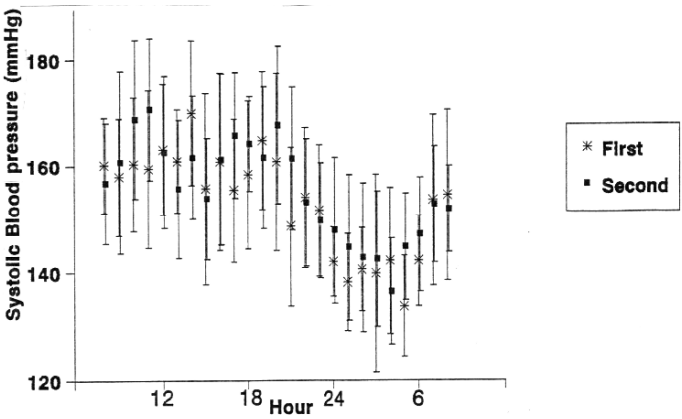


Figure 1. Mean values of ABPM data of 10 untreated patients with mild hypertension ands their SDs, recorded twice, one week in-between.

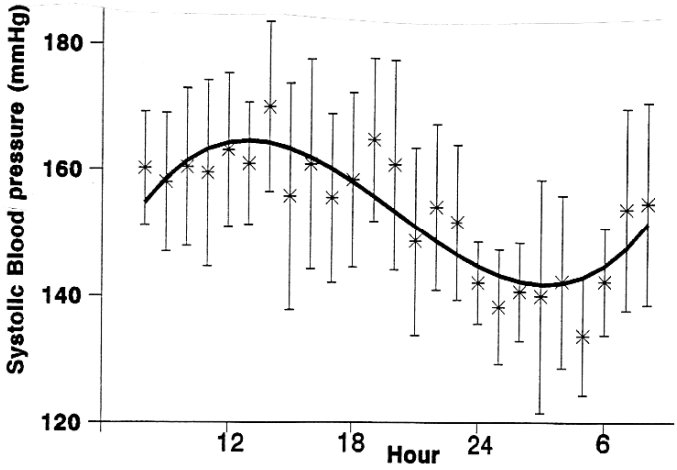


Figure 2. Polynome of corresponding ABPM recording (first one) from Figure 1, reflecting a clear circadian rhythm of systolic blood pressures.

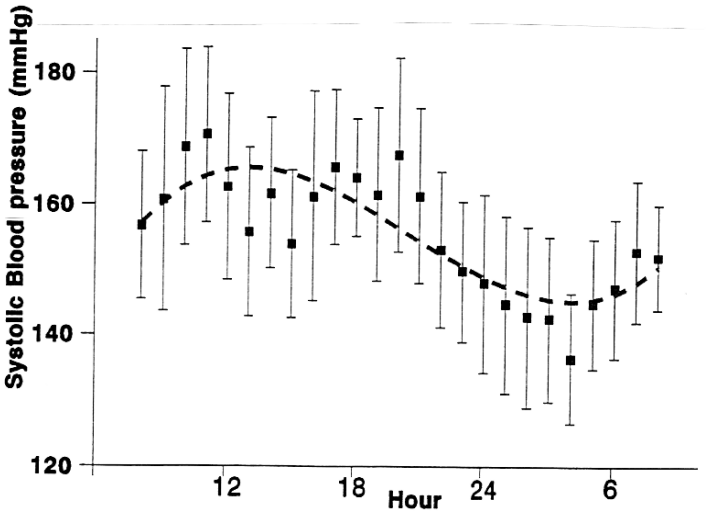


Figure 3. Polynome of corresponding ABPM recording (second one) from Figure 1, again reflecting a clear circadian rhythm of systolic blood pressures.

Figure 1 shows mean values of ABPM of 10 untreated patients and their SDs, recorded twice, one week in-between. Obviously, there is an enormous variance in the data both between-subject and within-subject as demonstrated respectively by the large SDs and the considerable differences between means. Figures 2 and 3 give polynomes of corresponding data from figure 1, reflecting a clear circadian rhythm in systolic blood pressures. Figure 4 shows that the two polynomes are, obviously, very much similar. Within-subject tests for reproducibility are given in Table I. Duplicate SDs of means versus zero and versus grand mean were 15.9 and 7.2, while of polynomes they were only 1.86 (differences in Duplicate SDs significant at a $P<0.001$ level). Intra-class correlations (ρ_Is) of means versus zero and versus grand mean were 0.46 and 0.75, while of polynomes they were 0.986 (differences in levels of correlation significant at a $P<0.001$). Obviously, polynomes of ABPM data of means of populations produce significantly better reproducibility than do the actual data.

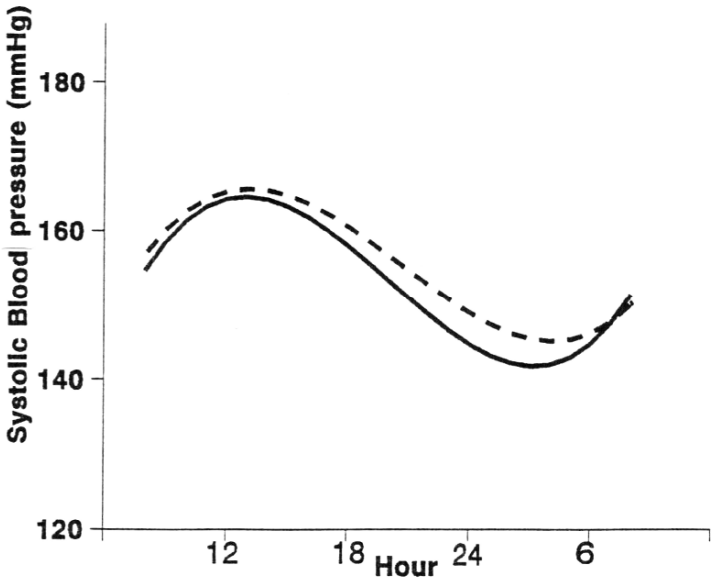


Figure 4. The two polynomes from Figures 2 and 3 are, obviously, very much similar.

Table 1.

24 hr ambulatory blood pressure measurements in a group of 10 patients with untreated mild hypertension tested twice: reproducibility of means of population (vs = versus)

	mean values variations vs zero	mean values variations vs grand mean	polynomes
Means (mm Hg) (test 1 / test 2)	153.1 / 155.4	153.1 / 155.4	-
SD (σ) (mm Hg) (test 1 / test 2)	21.9 / 21.1	15.7 / 13.8	-
95 % CIs[1] (mm Hg) (test 1 / test 2)	139.4-166.8/142.2-168.6	143.3-163.9/146.8-164.0	-
Differences between means (SD, σ) (mm Hg)	-2.4 (22.4)	-2.3 (10.5)	-
P values differences between results tests 1 and 2	0.61	0.51	0.44
Duplicate SDs[2] (mm Hg)	15.9	7.2	1.86
Relative Duplicate SDs[3] (%)	66	31	7
Intra-class correlations[4] (ρ_Is)	0.46	0.75	0.986
95 % CIs	0.35-0.55	0.26-0.93	0.972-0.999
Proportion total variance responsible for between-patient variance (%)	46	75	99
95 % CIs (%)	35-55	26-93	97-100

[1] CIs = confidence intervals.

[2] Duplicate SDs calculated according to Duplicate $SD = \sqrt{\frac{\sum (x_1 - x_2)^2}{2n}}$, where x_1 and x_2 are individual data during 1st and 2nd test, and n=240 (10 times 24 duplicate observations) .

[3] Calculated as 100%x[Duplicate SD/ (overall mean – 130 mm Hg)].

[4] Intra-class correlations (ρ_I s) calculated according to

$$\rho_1 = \frac{\sigma^2 \overline{x}_1 - \sigma^2 \overline{x}_2 / \overline{x}_1}{\sigma^2 \overline{x}_1}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the means of the 240 values during test 1 and test 2 respectively, and $\sigma^2 \overline{X}_2 / \overline{X}_1$ is the variance of $\overline{X}_2$ given $\overline{X}_1$, and

$$\sigma^2 \frac{\overline{x}_2}{\overline{x}_1} = \sigma^2 \overline{x}_1 - \frac{(\overline{x}_1 - \overline{x}_2)^2}{4}$$

Polynomes are, obviously, very much similar. Within-subject tests for reproducibility are given in Table 1. Duplicate SDs of means versus zero and versus grand mean were 15.9 and 7.2, while of polynomes they were only 1.86 (differences in Duplicate SDs significant at a $P<0.001$ level). Intra-class correlations (ρs) of means versus zero and versus grand mean were 0.46 and 0.75, while of polynomes they were 0.986 (differences in levels of correlation significant at a $P<0.001$). Obviously, polynomes of ABPM data of means of populations produce significantly better reproducibility, than do the actual data.

REPRODUCIBILITY OF INDIVIDUAL DATA

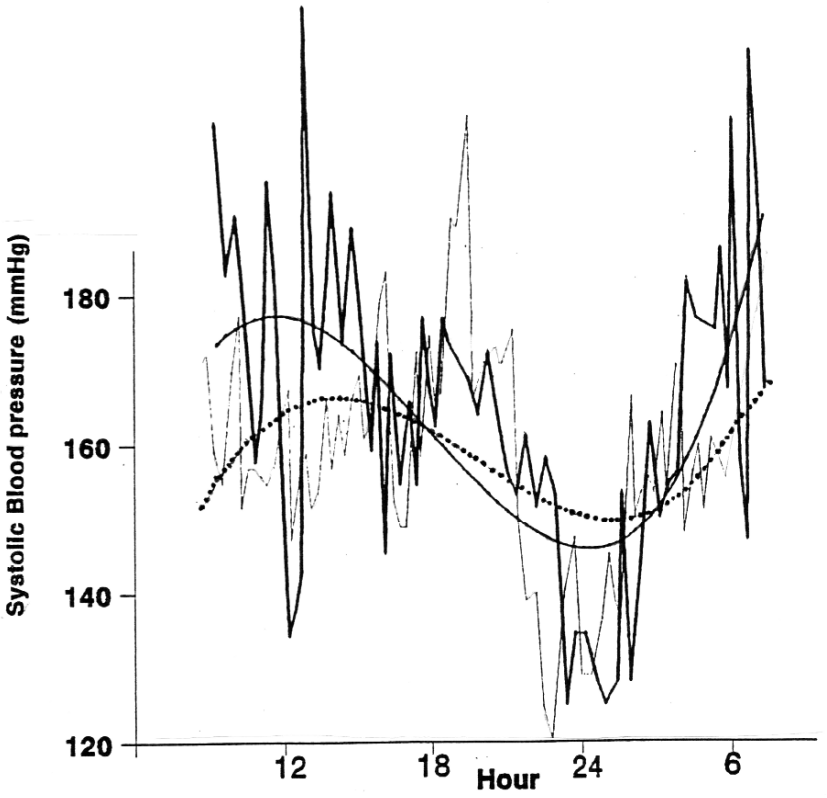


Figure 5. Individual data from patient 1 (Table 2) during first ABPM (fat line) and second ABPM recording (thin line). The corresponding polynomes of the two recordings (continuous and dotted curves respectively) are somewhat more different from each other than are the differences between the group data polynomes (Figure 4). Yet, they offer much better similarity than do the actual data.

Table 2.

24 hr ambulatory blood pressure measurements in 10 patients with untreated mild hypertension tested twice: reproducibility of individual data

Patient	mean (mm Hg) test 1 / test 2	SD (mm Hg) test 1 / test 2	Duplicate SDs (mm Hg) [1] raw data	polynomes	Intraclass Correlations[2] raw data	polynomes
1	160 / 157	14 / 18	17.7	2.1	0.07	0.58
2	158 / 161	17 / 27	17.6	9.0	0.27	0.53
3	160 / 169	20 / 29	19.7	2.6	-0.23	0.03
4	159 / 171	23 / 21	19.1	7.2	0.11	0.29
5	163 / 163	19 / 23	19.7	9.9	0.10	0.20
6	161 / 156	15 / 20	21.4	6.4	0.03	0.10
7	170 / 162	21 / 18	10.1	8.2	0.57	0.70
8	156 / 154	28 / 18	6.3	6.7	0.26	0.24
9	161 / 161	26 / 25	18.2	13.5	0.60	0.81
10	155 / 166	21 / 19	11.9	6.6	0.53	0.96
pooled data	153.1 / 155.4	21.9 / 21.1	16.2(5.0)[3]	7.2(3.3)	0.26(0.26)	0.42(0.34)
			___ P < 0.001___		___ P = 0.009____	

[1] Duplicate SDs calculated according to $SD = \sqrt{\frac{\sum (x_1 - x_2)^2}{2n}}$, where x_1 and x_2 are individual data during 1st and 2nd test, and n= 24 (24 duplicate observations per patient).

[2] Intra-class correlations (ρs) calculated according to

$$\rho_1 = \frac{\sigma^2 \overline{x}_1 - \sigma^2 \overline{x}_2 / \overline{x}_1}{\sigma^2 \overline{x}_1}$$

where $\overline{x}_1$ and $\overline{x}_2$ are the means of the 24 values during test 1 and test 2 respectively, and $\sigma^2 \overline{x}_2 / \overline{x}_1$ is the variance of $\overline{x}_2$ given $\overline{x}_1$, and

$$\sigma^2 \frac{\overline{x}_2}{\overline{x}_1} = \sigma^2 \overline{x}_1 - \frac{(\overline{x}_1 - \overline{x}_2)^2}{4}$$

[3] SDs between the brackets.

Figure 5 gives an example of the individual data of patient no. 1 during the first and second test and also shows his corresponding polynomes of test 1 and test 2. Although, again, there is enormous variability in the data, the polynomes have rather similar patterns. Table 2 gives an overview of assessments of reproducibility for each patient separately. Duplicate SDs of raw data were generally more than twice the size of those of the polynome*s*, while intraclass correlations of the actual

data were accordingly generally almost half the size of those of the polynomes with median values of 0.26 and 0.38 and ranges between –0.23 and 0.60 and between 0.03 and 0.96 respectively. Pooled differences were highly significant both for the Duplicate SDs, and for the intraclass correlations ($P < 0.001$ and $P = 0.009$ respectively, Table 2).

4. DISCUSSION

In this chapter we demonstrate that ABPM systolic blood pressures in untreated mildly hypertensive patients can be readily assessed by polynomial analysis and that this approach unlike the actual data analysis is highly reproducible. Similar results were obtained when instead of systolic blood pressures diastolic or mean pressures were analyzed. It may be argued from a mathematical-statistical point of view that the better reproducibility is a direct consequence of the procedure where variability is reduced by taking means of a population rather than individual values. However, when we compared polynomial and actual data for each subject separately, although the overall level of reproducibility fell, the former approach still performed better than did the latter. This indicates that the better reproducibility may at least in part be connected with mechanisms other than the mathematical necessity of reducing variability by taking the polynomial modeling of the actual data. Particularly, polynomes may be better reproducible, because they are a better estimate of the circadian rhythm of blood pressure than the actual data, which are of course influenced by a variety of exogenous factors including daily activities, meals and breaks, psychological effects. A polynome would be a more accurate estimate of the true endogenous circadian rhythm, where the mathematical procedure takes care that exogenous factors are largely removed. This would explain the high reproducibility not only of polynomial analyses of population data but also of individual patient data.

Polynomial analysis has been validated in chronobiology, as a reproducible method for the study of circadian rhythms in normotensive subjects, and is, actually, routinely used for that purpose in the Department of Chronobiology of our academic hospital.[11,12] So far, however, it has received little attention in the clinical assessment of patients with hypertension. The current chapter suggests, that the method would be a reliable instrument for that purpose.

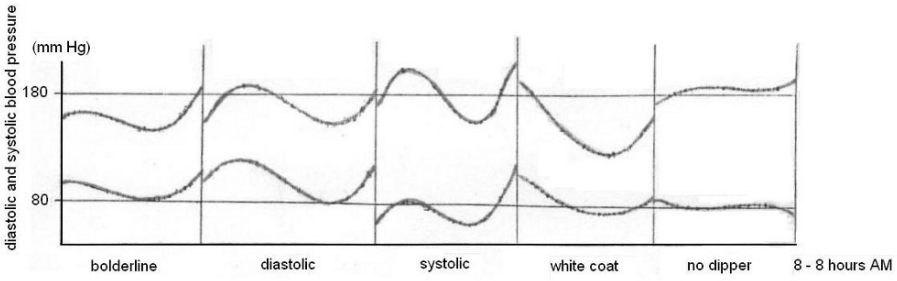


Figure 6. Polynomial analysis can be used to identify circadian patterns of blood pressure in individual patients.

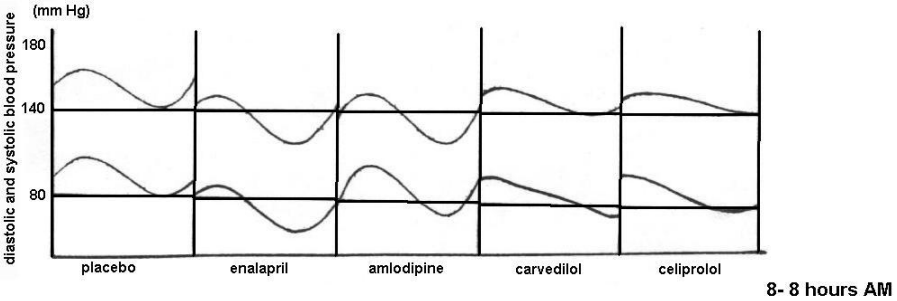


Figure 7. Polynomial analysis can be used to study the effects of antihypertensive drugs in groups of patients.

Polynomial analysis, could, e.g., be used to identify circadian patterns of blood pressure in individual patients. Figure 6 gives an example of 5 such patterns readily demonstrable by polynomes. These polynomes were drawn from ABPM data from our own outpatient clinic database. Figure 7 gives another example of how polynomes can be helpful in clinical assessments. The polynomes present the mean results of a recent study by our group, comparing the short term effects of different blood pressure reducing agents in mildly hypertensive patients (n=10).[6] All of the polynomes were within 95% CIs of the mean data of our samples. Differences between the data in this study, as assessed by 2-way analysis of variance, established that on enalapril, and amlodipine, unlike beta-blockers carvedilol and celiprolol, time effect was a major source of variability. The polynomes visualized that this was so, because beta-blockers did not reduce night-time blood pressures. So, polynomial analysis was helpful in interpreting the results of this study.

Polynomial analysis of ABPM data, unlike actual data analysis, is highly reproducible in patients with mild hypertension, and this is so not only with population means but also with individual data. It is, therefore, a valid approach for the clinical assessment of hypertensive patients, and may, thus, be helpful for a variety of purposes, e.g., for identifying circadian patterns of blood pressure in individual patients, and for the study of antihypertensive drugs in groups of patients. The goodness of fit of polynomial models estimated by levels of correlation between observed and modelled data, is very good, and sometimes even better than the real sine-like function derived from the Fourier analysis. Particularly, the regression lines of the 4th and 7th order generally provide the best fit for typical sinusoidal patterns. In the above example the 7th order polynome provided a slightly better fit than did the 4th order polynome.

5. CONCLUSIONS

Polynomial analysis is an extension of simple linear regression, where a power model is used to allow for the existence of a systematic, though not linear, dependence of the independent y variable on the dependent x variable, often a time variable. Particularly, fourth and seventh order polynomes are adequate to assess sinusoidal relationships, like circadian rhythms of hemodynamic and hormonal estimators.

6. REFERENCES

1. Owens P, Lyons S, O'Brien E. Ambulatory blood pressure in hypertensive population: patterns and prevalence of hypertensive subforms. J Hypertens 1998; 16: 1735-45.
2. Zanchetti A. Twenty-four-hour ambulatory blood pressure evaluation of antihypertensive agents. J Hypertens 1997; 15: S21-5.
3. Omboni S, Parati G, Palatini P, Vanasia A, Muiesan ML, Cuspidi C, Mancia G. Reproducibility and clinical value of nocturnal hypotension: prospective evidence from the SAMPLE study. J Hypertens 1998; 16: 733 - 8.
4. Bleniaszewski L, Staessen JA, Byttebier G, De Leeuw PW, Van Hedent T, Fagard R. Trough-to-peak versus surface ration in the assessment of antihypertensive agents. Blood Press 1997; 4: 350-7.
5. Van de Luit L, Van der Meulen J, Cleophas TJ, Zwinderman AH. Amplified amplitudes of circadian rhythms and nighttime hypotension in patients with chronic fatigue syndrome; improvement by inopamil but not by melatonin. Eur J Intern Med 1998; 9: 99-103.
6. Van de Luit L, Cleophas TJ, Van der Meulen J, Zwinderman AH. Nighttime hypotension in mildly hypertensive patients prevented by beta-blockers but not by ACE-inhibitors or calcium channel blockers. Eur J Intern Med 1998; 9: 251-6.
7. O'Brien E, Atkins N, Staessen J. State of the market, a review of ambulatory blood pressure-monitoring devices. Hypertension 1995; 26: 835-42.

8. Hays WL. Curvilinear regression. In: Hays WL, Statistics, Holt, Rinehart and Winston, Inc, Chicago, 4th edition, 1988, pp 698-716.
9. SPSS. Statistical Software. Professional Statistics. Chicago, Ill, 2002.
10. Harvard Graphics-3. Statistical Software. Boston MA, Harvard, Inc, 2001.
11. Scheidel B, Lemmer B, Blume H. Influence of time of day on pharmacokinetics and hemodynamic effects of beta-blockers. In: Clinical Chronopharmacology. Munich, Germany: Zuckschwerdt Verlag, 1990; vol 6: 75-9.
12. Lemmer B, Scheidel B, Behne S. Chronopharmacokinetics and chronopharmacodynamics of cardiovascular active drugs: propranolol, organic nitrates, nifedipine. Ann NY Acad Sci 1991; 618: 166-71.

CHAPTER 16

LOGISTIC AND COX REGRESSION, MARKOW MODELS, LAPLACE TRANSFORMATIONS

1. INTRODUCTION

Data modeling can be applied for improving precision of clinical studies. Multiple regression modeling is often used for that purpose. Relevant papers on this topic have recently been published.[1-7] Although multiple regression modeling, generally, does not influence the magnitude of the treatment effect versus control, it may reduce overall variances in the treatment comparison and thus increase sensitivity or power of statistical testing. It tries to fit experimental data in a mathematical model, and, subsequently, tests how far distant the data are from the model. A statistically significant correlation indicates that the data are closer to the model than will happen with random sampling. The very model-principle is at the same time its largest limitation: biological processes are full of variations and will not allow for a perfect fit. In addition, the decision about the appropriate model is not well founded on statistical arguments. The current study assesses uncertainties and risks of misinterpretations commonly encountered with regression analyses and rarely communicated in research papers. Simple regression models and real data examples are used for assessment.

2. LINEAR REGRESSION

Multiple linear regression for increasing precision of clinical trials assumes that a covariate like a baseline characteristic of the patients is an independent determinant of the treatment efficacy, and that the best fit for the treatment and control data is given by two separate regression lines with identical regression coefficients. The assumption may be too strong, and introduce important bias in the interpretation of the data, even if the variable seems to fit the model.

As an example is again taken the Regression Growth Evaluation Statin Study (REGRESS)[8] , a randomized parallel-group trial comparing placebo and pravastatin treatment in 434 and 438 patients, respectively. Primary endpoint was change in coronary artery diameter, secondary endpoint change in LDL (low density lipoprotein) cholesterol, as measured before and after two years of treatment. The average decreases of LDL cholesterol are

statin:	1.23 (standard deviation (SD) 0.68) mmol/l
placebo:	-0.04 (SD 0.59) mmol/l

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 199–220.

Obviously, LDL decrease varies considerably in both treatment groups but, on average, treatment efficacy can be quantified as 1.23 – (-0.04) = 1.27 mmol/l. Since the patients in the two parallel groups are independent of each other, the standard error (SE) of this estimate equals

$$\sqrt{\frac{0.682}{438}+\frac{0.592}{434}} = 0.043 \text{ mmol/l.}$$

The same results can be obtained by drawing the best fit for the data in the form of a regression line according to the equation:

$$y = a + b\,x,$$

where

y = the dependent variable representing the LDL cholesterol decrease of the patients,

x = the independent variable representing treatment modality, 1 if a patient receives statin, and 0 if placebo.

The term a is the intercept of the regression line with the y-axis and b is the regression coefficient (= direction coefficient of the regression line) which must be estimated.

Figure 1 gives the linear regression line in graph. It yields an estimate of b of 1.27 with SE 0.043; hence, completely equal to the above analysis.

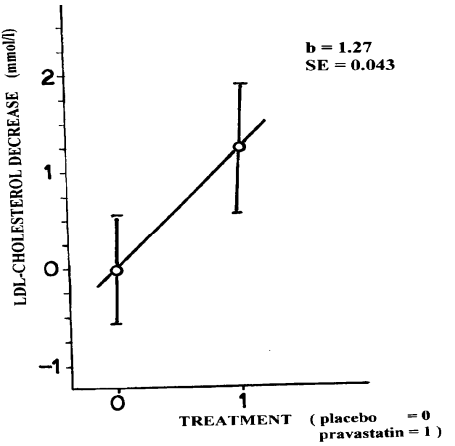


Figure 1. The linear regression line is illustrated (b = regression coefficient, SE = standard error).

We wish to adjust these data for baseline LDL cholesterol. First, we draw a scatter plot of the individual baseline LDL cholesterol values and LDL cholesterol decreases (Figure 2 (1)). Both on placebo and on active treatment a positive linear correlationship is suggested between LDL cholesterol decrease and baseline LDL

cholesterol: the larger the baseline LDL cholesterol the better the LDL cholesterol-decrease. Figure 2 (2) shows that the overall linear correlation between these two variables is, indeed, significant at $p<0.001$. Baseline LDL cholesterol is thus an independent determinant of LDL cholesterol decrease.

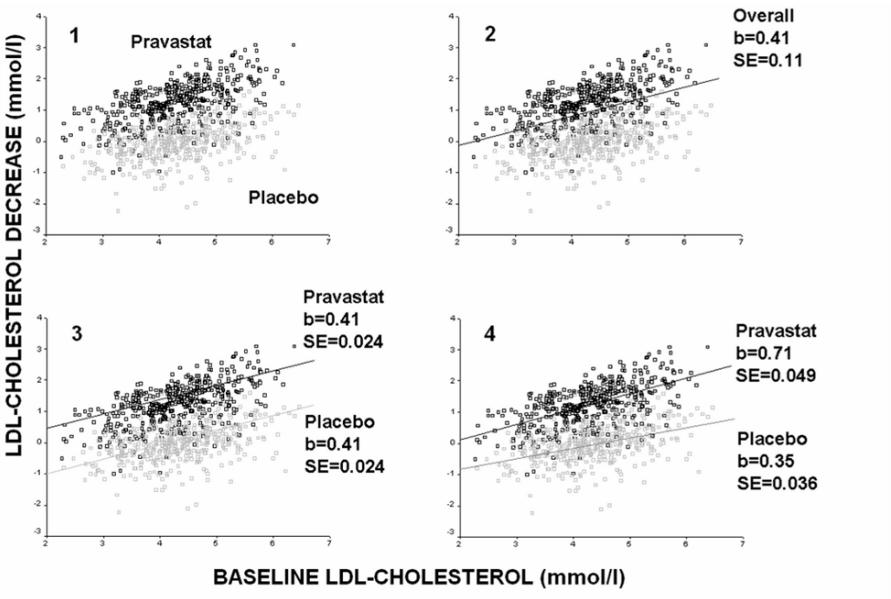


Figure 2. 1. Both on placebo and on active treatment there seems to be a positive correlation between LDL cholesterol decrease and baseline LDL cholesterol: the larger the baseline LDL cholesterol the better the LDL cholesterol decrease. 2. Overall correlation is significant at $p<0.001$, baseline LDL cholesterol is thus an independent determinant of LDL cholesterol decrease. 3. The multiple linear regression model assesses whether the data are significantly closer to two regression lines with identical regression coefficients (= direction coefficients) than compatible with random sampling. 4. The separately calculated regression lines are not parallel (regression coefficients 0.71 (SE 0.049, $p<0.0001$) and 0.35 (0.036, $p<0.0001$, difference in slope 0.36 (SE 0.06, $p<0.0001$)); (b = regression coefficient, SE = standard error).

To test whether this significant independence remains after adding the variable treatment modality to the regression, we use the following (multiple) linear regression model:

$$y = a + b_1 x_1 + b_2 x_2$$

where

y = the dependent variable representing the LDL cholesterol decrease of the patients,

x_1 = the independent variable representing treatment modality, 1 if a patient receives statin, and 0 if placebo,
x_2 = a second independent variable, baseline LDL cholesterol.

An Excel data file, entered into SPSS Statistical Software, produces the following results:

$b_2 = 0.41$ (SE = 0.024, $p<0.0001$),
$b_1 = 1.27$ (SE = 0.037, $p<0.0001$).

Figure 2 (3) shows how the model works. It assesses whether the data are significantly closer to two regression lines with identical regression coefficients (= direction coefficients) than compatible with random sampling.
With placebo ($x_1 = 0$) the best fit for the data is given by the formula

$$y = a + b_2 x_2,$$

With pravastatin ($x_1 = 1$) the best fit for the data is given by the formula

$$y = a + b_1 + b_2 x_2.$$

The estimated treatment effect, b1, is 1.27, the same as in the simple linear regression from Figure 1, but its SE is lowered from 0.043 to 0.037. This means that, indeed, increased precision has been obtained by the multiple regression modeling. The difference between the two regression lines represents the treatment efficacy of pravastatin versus placebo: for each point on the x-axis (baseline LDL cholesterol) the average LDL cholesterol decrease is 1.27 mmol/l larger in the statin (grey) group than in the placebo (black) group. The positive linear correlation between LDL cholesterol decrease and baseline LDL cholesterol (the larger the baseline LDL cholesterol the better the LDL cholesterol decrease) in either of the groups could be explained by a regression-to-the-mean-like-phenomenon: the patients scoring low the first time are more likely to score higher the second time vice versa. However, why should the best fit regression lines of the pravastatin data and of the placebo data produce exactly the same regression coefficients. In order to assess this question regression lines for either of the groups can be calculated separately. Figure 2 (4) shows the results. In contrast with the multiple linear regression lines, the separately calculated regression lines are not parallel. Their regression coefficients are 0.71 (SE = 0.049, $p<0.0001$) and 0.35 (SE = 0.036, $p<0.0001$). The difference in slope is significant with a difference in regression of 0.36 (SE = 0.06, $p<0.0001$). Obviously, there is no homogeneity of regression for the groups.

If the parallel regression lines from Figure 2 (3) are interpreted as a homogeneous regression-to-the-mean-like-phenomenon in either of the two treatment groups, then the appropriate implication will be that pravastatin's efficacy is independent of baseline LDL cholesterol. However, the true regression lines from Figure 2 (4) indicate that there is a significant difference in slope. This difference in slope can only be interpreted as a dependency of pravastatin's efficacy on baseline LDL cholesterol: the higher the baseline-cholesterol the better the efficacy of treatment.

In clinical terms, based on the multiple regression analysis all patients no matter their baseline LDL cholesterol would qualify for pravastatin treatment equally well, while based on the true regression lines patients would qualify better the higher their baseline LDL cholesterol.

3. LOGISTIC REGRESSION

LOGISTIC REGRESSION ANALYSIS FOR PREDICTING THE PROBABILITY OF AN EVENT

Th odds of an infarction is given by the equation

$$\text{odds infarct in a group} = \frac{\text{number of patients with infarct}}{\text{number of patients without}}$$

The odds of an infarction in a group is correlated with age, the older the patient the larger the odds

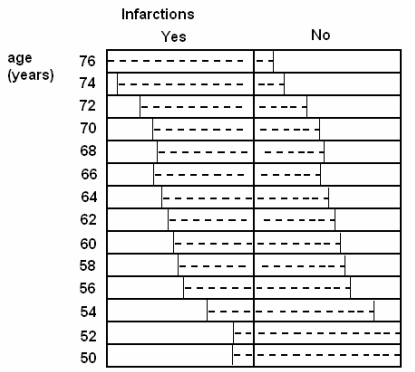


Figure 3. In a group of multiple ages the numbers of patients at risk of infarction is given by the dotted line

According to Figure 3 the odds of infarction is correlated with age, but we may ask how?

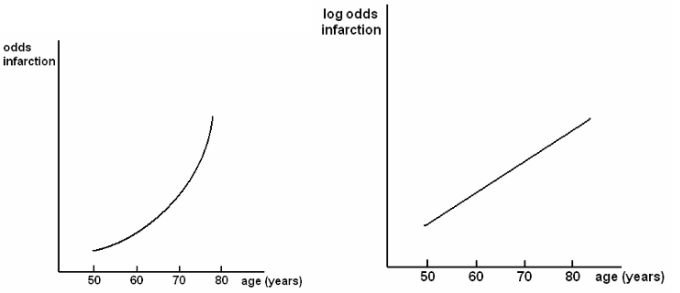


Figure 4. Relationships between the odds of infarction and age

According to Figure 4 the relationship is not linear, but after transformation of the odds values on the y-axis into log odds values the relationship is suddenly linear.
We will, therefore, transform the linear equation

$$y = a + bx$$

into a log linear equation (ln = natural logarithm)

$$\ln \text{odds} = a + b\,x \quad (x = \text{age})$$

Our group consists of 1000 subjects of different ages that have been observed for 10 years for myocardial infarctions. Using SPSS statistical software, we command binary logistic regression

dependent variable infarction yes / no (0 / 1)
independent variable age

The program produces a regression equation:

$$\ln \text{odds} = \ln \frac{\text{pts with infarctions}}{\text{pts without}} = a + bx$$

a = -9.2
b = 0.1 (SE = 0.04; $p < 0.05$)
The age is, thus, a significant determinant of odds infarction (which can be used as surrogate for risk of infarction).
Then, we can use the equation to predict the odds of infarction from a patient's age:

Ln odds55 years = -9.2 + 0.1 . 55 = -4.82265
odds = 0.008 = 8 / 1000

Ln odds75 years = -9.2 + 0.1 . 75 = -1.3635
odds = 0.256 = 256 / 1000

Odds of infarction can, of course, more reliably be predicted from multiple x-variables. As an example, 10,000 pts are followed for 10 years, while infarctions and baseline-characteristics are registered during that period.

dependent variable infarction yes/no

independent variables gender
predictors age
Bmi (body mass index)
systolic blood pressure
cholesterol
heart rate
diabetes
antihypertensives
previous heart infarct
smoker

The data are entered in SPSS, and it produces b-values (predictors of infarctions)

	b-values	p-value
1.Gender	0.6583	< 0.05
2.Age	0.1044	"
3.Bmi	-0.0405	"
4.Systolic blood pressure	0.0070	"
5.Cholesterol	0.0008	"
6.Heart rate	0.0053	"
7.Diabetes	1.2509	< 0.10
8.Antihypertensives	0.3175	< 0.050
9.Previous heart infarct	0.8659	< 0.10
10.Smoker	0.0234	< 0.05
a-value	-9.1935	"

It is decided to exclude predictors that have a p-value > 0.10.
The regression equation is used

"ln odds infarct = $a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots$"

to calculate the best predictable y-value from every single combination of x-values. For instance, for a subject

-Male (x_1)
-55 years of age (x_2)
-cholesterol 6.4 mmol/l (x_3)
-systolic blood pressure 165 mmHg (x_4)
-antihypertensives (x_5)
-dm (x_6)
-15 cigarettes / day (x_7)
-heart rate 85 beats / min (x_8)
-Bmi 28.7 (x_9)
-smoker (x_{10})

the calculated odds of having an infarction in the next 10 years is the following:

	b-values	x-values		
Gender	0.6583	. 1 (0 or 1)	=	0.6583
Age	0.1044	. 55	=	5.742
BMI	-0.0405	. 28.7	=	..
Blood pressure	0.0070	. 165	=	
Cholesterol	0.0008	. 6.4	=	
Heart rate	0.0053	. 85	=	
Diabetes	1.2509	. 1	=	
Antihypertensives	0.3175	. 1	=	

Previous heart inf	0.8659	.	0	=
Smoker	0.0234	.	15	=
a-value				= -9.1935 +
			Ln odds infarct	= -0.5522
			odds infarct	= 0.58 = 58/100

The odds is often interpreted as risk. However, the true risk is a bit smaller than the odds, and can be found by the equation

risk event = 1 / (1 + 1/odds)

If odds of infarction = 0.58, then the true risk of infarction = 0. 37.
The above methodology is currently an important way to determine, with limited health care sources, what individuals will be:
(1) operated.
(2) given expensive medications.
(3) given the assignment to be treated or not.
(4) given the "do not resuscitate sticker".
(5) etc.
We need a large data base to obtain accurate b-values. This logistic model for turning the information from predicting variables into probability of events in individual subjects is being widely used in medicine, and was, for example, the basis for the TIMI (Thrombolysis In Myocardial Infarction) prognostication risk score. However, not only in medicine, also in strategic management research, psychological tests like computer adapted tests, and many more fields it is increasingly observed (Table 1). With linear regression it is common to provide a measure of how well the model fits the data, and the squared correlation coefficient r^2 is mostly applied for that purpose. Unfortunately, no direct equivalent to r^2 exists for logistic, otherwise called loglinear, models. However, pseudo-R2 or R2-like measures for estimating the strength of association between predictor and event have been developed.

Table 1. Examples of predictive models where multiple logistic regression has been applied

	dependent variable (odds of event)	independent variables (predictors)
1. TIMI risk score[9]	odds of infarction	age, comorbidity, comedication, riskfactors
2. Car producer (Strategic Management Research)[10]	odds of successful car	cost, size, horse power, ancillary properties
3. Item response modeling (Rasch models for computer adapted tests)[11]	odds of correct answer to three questions of different difficulty	correct answer to three previous questions

LOGISTIC REGRESSION FOR EFFICACY DATA ANALYSIS

Logistic regression is often used for comparing proportions of responders to different treatments. As an example, we have two parallel groups treated with different treatment modalities.

	Responders	non-responders
New Treatment (group 1)	17 (A)	4 (B)
Control Treatment (group 2)	19 (C)	28 (D)

The odds of responding is given by A/B and C/D, and the odds ratio by $\frac{A/B}{C/D}$.

It has been well-established that no linear relationship exists between treatment modalities and odds of responding, but that there is a close-to-linear relationship between treatment modalities and the logarithm of the odds. The natural logarithm (ln) even better fits such assessments. And so, for the purpose of the logistic regression we assume that the usual linear regression formula

$$y = a+bx$$

is transformed into

$$\ln \text{odds} = a+bx,$$

where ln odds = the dependent variable,
x = the independent variable representing treatment modality, 1 if a patient receives new treatment , and 0 if control treatment.

The term a is the intercept of the regression line, and b is the regression coefficient (direction coefficient of the regression line).

Instead of ln odds = a +bx
the equation can also be described as

$$\text{odds} = e^{a+bx}$$
$$\text{odds}_{\text{new treatment}} = e^{a+b} \quad \text{because } x=1$$
$$\text{odds}_{\text{control treatment}} = e^{a} \quad \text{because } x=0$$
$$\text{odds ratio} = e^{a+b} / e^{a} = e^{b}$$

The new treatment is significantly different from the control treatment if the odds ratio of the two treatments is significantly different from 1. If b = 0, then the odds ratio = $e^0 = 1$, which means no difference between new and control treatment. If b is significantly > 0, then the odds ratio is significantly > 1, which means a significant difference between new and control treatment.

SPSS Statistical Software produces the best fit b and a for the data:

a = –1.95 (SE = 0.53)
b = 1.83 (SE = 0.63, p=0.004).

The estimated b is significantly different from 0 at p=0.004, and so we conclude that new and control treatment are significantly different from one another. A similar result could have been obtained by the usual chi-square test. However, the logistic model can adjust the results for relevant subgroups variables like age, gender, and concomitant illnesses. In our case, the data are divided into two age groups

	responders	non-responders	responders	non-responders
	> 50 years		<50 years	
Group 1	4	2	13	2
Group 2	9	16	10	12

The underlying assumptions are that the chance of response may differ between the subgroups, but that the odds ratio does not. SPSS Statistical Software calculates the best fit b- and a-values for data:

$a_{>50 \text{ years}}$ = –2.37 (SE = 0.65)
$a_{<50 \text{ years}}$ = -1.54 (SE = 0.59)
b =1.83 (SE = 0.67, p= 0.007)

The estimated b is significantly different from 0 also after age-adjustment. Figure 5 shows how the model works. Like with the linear regression model it assesses whether the data are closer to two regression lines with identical regression coefficients than compatible with random. However, why should the best fit regression lines of the different age groups produce exactly the same regression coefficients? Regression lines for either group can be calculated separately to answer this question. In contrast to the logistic regression lines the separately calculated regression lines are not parallel. Their regression coefficients are 1.27 (SE = 0.39, p<0.001) and 2.25 (SE = 0.48 p<0.001). The difference in slope is significant, with a difference in regression of 0.98 (SE = 0.60, p < 0.05). Obviously,

there is no parallelism between the groups. Younger patients not only respond better, but also benefit more from the new than from the control treatment. This latter mechanism of action is clinically very relevant but remains unobserved in the logistic regression analysis.

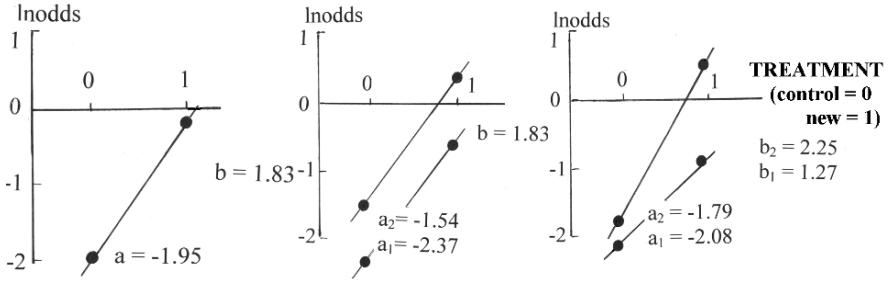


Figure 5. Left graph shows a linear correlation between ln odds of responding and treatment modalities (b=1.83, SE = 0.63, p=0.004). The logistic model (middle graph) assesses whether the data are closer to two regression lines with identical direction coefficients than compatible with random sampling. The separately calculated regression lines (right graph) are not parallel (regression coefficients 2.25 (SE = 0.38, p<0.001) and 1.27 (SE = 0.48, p<0.001), difference in slope 0.98 (SE = 0.60, p < 0.05); (b = regression coefficient, a = intercept, SE = standard error).

4. COX REGRESSION

Cox regression is based on the assumption that per time unit approximately the same percentage of subjects at risk will have an event, either deadly or not. This exponential model is suitable for mosquitos whose risk of death is determined by a single factor, i.e., the numbers of collisions, but less so for human beings whose deaths are, essentially, multicausal. Yet, it is widely applied for the comparison of two Kaplan-Meier curves in human beings. Figure 6 shows that after 1 day 50% is alive, while after the second day 25% is, etc.

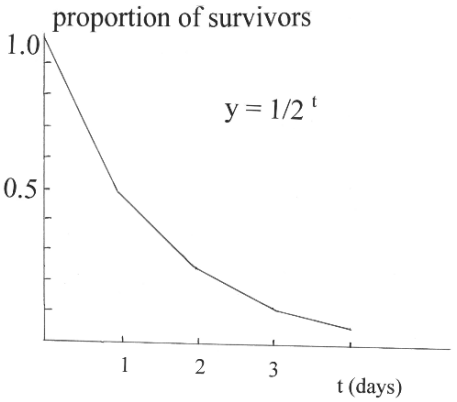


Figure 6. Hypothesized example of the exponential surviving pattern of mosquitos.

The formula for the proportion of survivors is given by:

proportion survivors = $1/2^{t} = 2^{-t}$

In true biology "e (= 2.71828)" instead of "2" better fits the observed data, while k is dependent on the species:

proportion survivors = e^{-kt}

The Cox regression formula for the comparison of exponential survival curves is given by:

proportion survivors = $e^{-kt\ -bx}$,
x = binary variable (only 0 or 1; 0 means treatment-1, and 1 means treatment-2),
b = regression coefficient,
proportion survivors treatment-1 = e^{-kt} because x = 0,
proportion survivors treatment-2 = $e^{-kt\ -\ b}$ because x = 1,
relative risk of surviving = $e^{-kt-b} / e^{-kt} = e^{-b}$,
relative risk of death = hazard ratio = e^{b}.

Figure 7 shows two Kaplan-Meier curves. Although an exponential pattern is hard to prove from the curves (or from their logarithmic transformations), the Cox model seems reasonable, and SPSS software is used to calculate the best b for the given data.

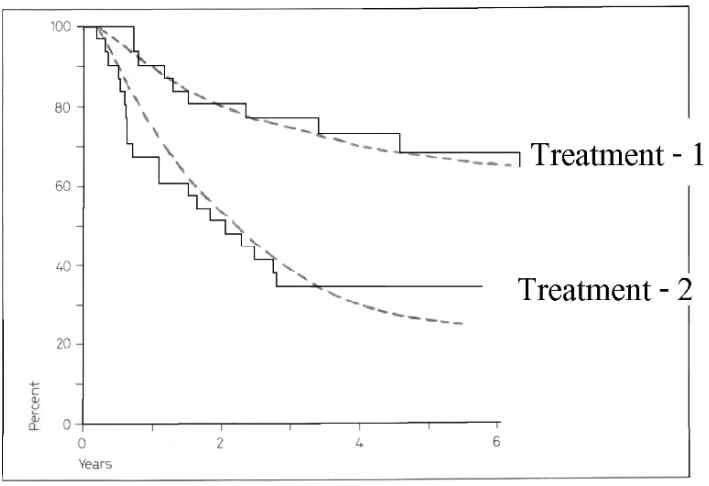


Figure 7. Two Kaplan-Meier curves estimating effect on survival of treatment 1 and 2 in two parallel groups of patients with malignancies (33 and 31 patients respectively). The dotted curves present the modeled curves produced by the Cox regression model.

If b is significantly larger than 0, the hazard ratio will be significantly larger than 1, and there will, thus, be a significant difference between treatment-1 and treatment-2. The following results are obtained:

b = 1.1 with a standard error of 0.41
hazard ratio = 3.0
p = 0.01 (t-test)

The Cox regression provides a p-value of 0.01, and, so, it is less sensitive than the traditional summary chi-square test (p-value of 0.002). However, the Cox model has the advantage that it enables to adjust the data for relevant prognostic factors like disease stage and presence of b-symptoms. The model is extended accordingly:

hazard ratio = $e^{b_1x_1+b_2x_2+b_3x_3}$
$x_1 = 0$ (treatment-1); $x_1 = 1$ (treatment-2)
$x_2 = 0$ (disease stage I-III); $x_2 = 1$ (disease stage IV)
$x_3 = 0$ (A symptoms); $x_3 = 1$ (B symptoms)

The test for multicollinearity is negative (Pearson correlation coefficient between disease stage and B symptoms < 0.85), and, so, the model is deemed appropriate. SPSS produces the following result:

$b_1 = 1.10$ with a standard error of 0.41
$b_2 = 1.38$ " " " 0.55
$b_3 = 1.74$ " " " 0.69
unadjusted hazard ratio = 3.0
adjusted hazard ratio = 68.0

Treatment-2 after adjustment for advanced disease and b-symptoms raises a 68 higher mortality than treatment-1 without adjustments. This Cox regression analysis, despite prior examination of the appropriateness of the model, is hardly adequate for at least three reasons. First, the method is less sensitive than the usual chi-square summary test, probably because the regression does not fit the data well enough. Second, Cox regression tests the null-hypothesis that treatment-2 is not significantly different from the treatment-1, and it assumes for that purpose that the hazard ratio is constant over time. Figure 5 gives the modeled treatment-curves (dotted curves), in addition to the true treatment-curves. It can be observed in the modeled curve that few patients died in the first 8 months, while, in reality, 34% of the patients in group 2 died, probably, due to the toxicity of the treatment-2. Also it can be observed in the modeled curves that patients continued to die after 2 1/2 years, while, in reality, they stopped dying in group 2, because they actually went into a complete remission. Obviously, this Cox regression analysis gives rise to some serious misinterpretations of the data. Third, a final problem with the above Cox analysis is raised by the adjustment-procedure. An adjusted hazard ratio as large as 68 is clinically unrealistic. Probably, the true adjusted hazard ratio is less than 10. From a clinical point of view, the x_2 and x_3 variables must be strongly dependent on one another as they are actually different measures for estimating the same. And so, despite the negative test for multicollinearity, they should not have been included in the model.

Note: Cox regression can be used for other exponential time relationships like pharmacokinetic data. Limitations similar to ones described above apply to such analyses.

5. MARKOW MODELS

Regression models are only valid within the range of the x-values. Markow modeling goes one step further, and aims at predicting outside the range of x-values. Like with Cox regression it assumes an exponential-pattern in the data which may be a strong assumption.

As an example, in patients with diabetes mellitus type II, sulfonureas are highly efficacious, but they will, eventually, induce beta-cell failure. Beta-cell failure is sometimes defined as a fasting plasma glucose >7.0 mmol/l. The question is, does the severity of diabetes and / or the potency of the sulfonurea-compound influence the induction of beta-cell failure?

This was studied in 500 patients with diabetes type II:

at time 0 year	0 / 500 patients	had beta-cell failure
at time 1 year	50 / 500 patients (=10%)	had beta-cell failure.

As after 1 year 90% had no beta-cell failure, it is appropriate according to the Markow model to extrapolate:

after 2 years 90% x 90% = 81% no beta-cell failure
after 3 years 90% x 90% x 90% = 73% no beta-cell failure
after 6.7 years = 50% no beta-cell failure.

A second question was, does the severity of diabetes mellitus type II influence induction of beta-cell failure. A cut-off level for severity often applied is a fasting plasma glucose > 10 mmol/l. According to the Markow modeling approach the question can be answered as follows:

250 patients had fasting plasma glucose < 10 mmol/l at diagnosis (Group-1)
250 patients had fasting plasma glucose > 10 mmol/l at diagnosis (Group-2)

If after 1 year sulfonureas (su) treatment 10 / 250 of the patients from Group -1 had b-cell failure, and 40 / 250 of the patients from Group-2, which is significantly different by $p<0.01$, then we can again extrapolate:

In Group-1 it takes 12 years before 50% of the patients develop beta-cell failure.
In Group-2 it takes 2 years before 4% of the patients develop beta-cell failure.

The next question is, does potency of su-compound influence induction of b-cell failure?

250 patients started on amaryl (potent sulfonurea) at diagnosis (Group-A)
250 patients started on artosin (non-potent sulfonurea) at diagnosis (Group-B)

If after 1 year 25 / 250 of Group-A had beta-cell failure, and 25 / 250 of the group-B, it is appropriate according to the Markow model to conclude that a non-potent does not prevent beta-cell failure. Note Markow modeling is highly speculative, because nature does not routinely follow mathematical models.

6. REGRESSION-ANALYSIS WITH LAPLACE TRANSFORMATIONS

There is an increasing trend towards the use of non linear mixed effect models (commonly called population pharmacokinetics and pharmacodynamics) for describing the pharmacokinetics and pharmacodynamics of drugs in humans. The term mixed effect models refers to the random effect statistical regression model applied. These models allow for sparse sampling and at the same time can account for multiple effect associated variables and even account for errors in sampling.[12-14] These new modelling approaches are increasingly becoming a very important part

of the drug approval process. They routinely make use of multi-exponential models, according to equations like for example the one underneath:

$f(t) = D/V\ (e^{-at} + e^{-bt} + e^{-ct} \ldots.)$

D = dose drug
V = volume of distribution
a = elimination constant compartment 1
b = elimination constant " 2
c = elimination " " 3
t = time

As logarithmic transformations only allow for mono-exponential equations, generally Laplace transformations, based on second differentiations, are used:

$f(t) = C(t) = C(0)\ (e^{-at} + e^{-bt})$

C(t) = concentration at time t, C(0) = concentration at time 0.

is transformed into

$F(s) = C(0) / (s+a)(s+b)$

s = unit Laplace-functions = unit amount-of-drug / time
(time thus disappears from the equation).
The final data are then transformed back to their initial equations.

The advantage of the exponential modeling in pharmacokinetics that it is very easy to calculate the keystone pharmacokinetic parameters according to which compounds are currently registered: plasma-half-life, volume of distribution, plasma-clearance rate etc.
Exponential pharmacokinetic models assume first order kinetics, and it may be true that many drugs at the therapeutically given concentrations would follow first order kinetics. However, zero order patterns are followed for example by ethyl-alcohol, acetyl-salicylic-acid, and by any drug at higher dosages, while second order elimination-patterns are followed for example by drugs that are hydrolyzed or conjugated before excretion.[15] The simplest equations and curves for zero, first, and second order kinetics are given (Figure 8).

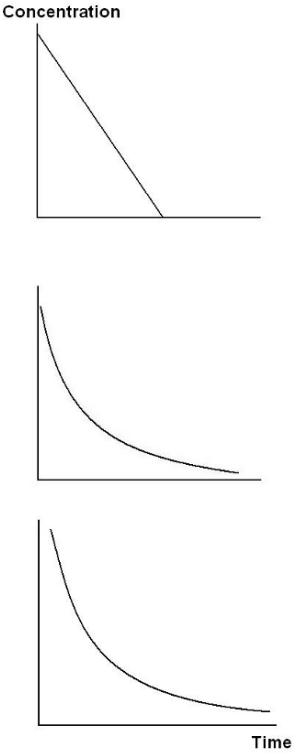


Fig. 8. Examples of hypothesized time-concentration curve following zero-, first-, and second order pharmacokinetics.

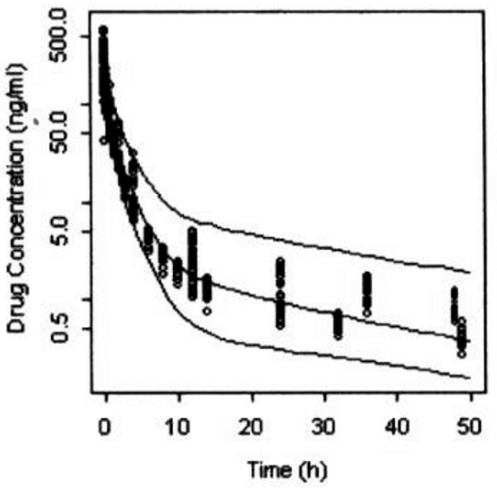


Fig. 9. Example of an exponentially modeled time-concentration relationship with wide 95% confidence intervals.

1. Zero order $C(t) = C(0) - kt$ linear pattern
2. First order $C(t) = C(0) \cdot e^{-kt}$ exponential pattern
3. Second order $1 / C(t) = 1 / C(0) - kt$ hyperbolic pattern

k = elimination constant

As shown in the example of Figure 9, there may be a wide spread in the data of a pharmacokinetic study. The 95% confidence intervals calculated with the NON-MEM software[12], which uses the Laplace transformations, assumes a first order pharmacokinetic. In fact both a zero and a second order pattern provided a better fit in this example. However, a problem with either of them is, that it is impossible to derive plasma-half-life and other pharmacokinetic parameters from them. As can be observed only in equation 2 plasma-half-life is not dependent on C(0). With equations 1 and 3 we have many plasma-half-lifes, with equation 2 we have only one. This is a very elegant property of first order kinetics, but it should not mean that the best fit data models are sacrificed for the purpose of unreliable pharmacokinetic parameters. A second problem with the Laplace models is that they assume independence of confounders. In pharmacology confounders like gender, age, body mass, renal function, notoriously interact with the treatment modalities.

7. DISCUSSION

In randomized controlled trials regression analysis of possibly confounding variables is, traditionally, not emphasized, because the randomization ensures that such variables are equally distributed among the treatment groups. Also, regression analysis tests correlations rather than causal relationships. In testing the data from clinical trials we are mainly interested in causal relationships. When such assessments were statistically analyzed through correlation analyses, we would probably be less convinced of a causal relationship than we are while using prospective hypothesis testing. In the past few years, however, regression analyses have increasingly entered the stage of primary data analysis. E.g., of 28 randomized controlled trials published in the Lancet in the 2003 Volume 362, 20 (71%) used regression models for primary data analysis, including linear regression twice, logistic regression five times, and Cox regression twelve times.

Obviously, regression analyses are increasingly used for the primary data analysis of clinical trials. The current paper assesses problems of this new development. More uncertainties are added to the data in the form of subjective judgments and uncertainty about the appropriate transformation of the data. Regression analyses may also give rise to serious misinterpretations of the data:

1. The assumption that baseline characteristics are independent of treatment efficacies may be wrong.
2. Sensitivity of testing is jeopardized if the models do not fit the data well enough.
3. Relevant clinical phenomena like unexpected toxicity effects and complete remissions can go unobserved.
4. The inclusion of multiple variables in regression models raises the risk of clinically unrealistic results.

Markow modeling is an exponential regression model like Cox regression that aims at predicting outside the range of observed observations. It, is, therefore even more at risk of unrealistic results. As an example, many suggestions from the famous Framingham studies are based on Markow modeling. Current trials and observations confirm that some of these are true, some are not. Regression modeling, although a very good tool for exploratory research, is not adequately reliable for randomized clinical trials. This is, of course, different with exploratory research like observational studies. E.g., a cohort of postmenopausal women is assessed for exploratory purposes. The main question is: what are the determinants of endometrial cancer in this category op females. Logistic regression is excellent for the purpose of this exploratory research. The following logistic model is used:

y-variabele = ln odds endometrial cancer
x_1 = estrogene consumption short term
x_2 = estrogene consumption long term
x_3 = low fertility index

x_4 = obesity
x_5 = hypertension
x_6 = early menopause

lnodds endometrial cancer = a + b_1 estrogene data + b_2....+ b_6 early menopause data

The odds ratios for different x-variables are defined, e.g., for:
x_1 = chance cancer in consumers of estrogene / non-consumers
x_3 = chance cancer in patients with low fertility / their counterparts
x_4 = chance cancer in obese patients / their counterparts etc.

risk factors	regression coefficient(b)	standard error	p-value	odds ratio (e^b)
1.estrogenes short	1.37	0.24	<0.0001	3.9
2.estrogenes long	2.60	0.25	<0.0001	13.5
3.low fertility	0.81	0.21	0.0001	2.2
4.obesity	0.50	0.25	0.04	1.6
5.hypertension	0.42	0.21	0.05	1.5
6.early menopause	0.53	0.53	ns	1.7

The data are entered in the software program, which provides us with the best fit b-values. The model not only shows a greatly increased risk of cancer in several categories, but also allows us to consider that the chance of cancer if patients consume estrogens, suffer from low fertility, obesity, and hypertension might have an increased risk as large as = $e^{b2+b3+b4+b5}$ = 75.9 = 76 fold. This huge chance is, of course, clinically unrealistic! We must take into account that some of these variables must be heavily correlated with one another, and the results are, therefore, largely inflated. In conclusion, regression modeling is an adequate tool for exploratory research, the conclusions of which must be interpreted with caution, although they often provide scientifically highly interesting questions. Such questions are, then, a sound basis for confirmation by prospective randomized research. Regression modelling is not adequately reliable for the analysis of the primary data of randomized controlled trials. Of course, regression analysis is also fully in place for the exploratory post-hoc analyses of randomized controlled trials (chapters 16, 17, and 29).

8. CONCLUSIONS

Data modeling can be applied for improving precision of clinical studies. Multiple regression modeling is increasingly used for that purpose. The objective of this chapter was to assess uncertainties and risks of misinterpretations commonly encountered with regression analyses and rarely communicated in research papers.
Regression analyses add uncertainties to the data in the form of subjective judgments and uncertainty about the appropriate transformation of the data. Additional flaws include: (1) the assumption that baseline characteristics are

independent of treatment efficacies; (2) the loss of sensitivity of testing if the models do not fit the data well enough; (3) the risk that clinical phenomena like toxicity effects and complete remissions go unobserved; (4) the risk of clinically unrealistic results if multiple variables are included. Regression analyses, although a very good tool for exploratory research, are not adequately reliable for randomized controlled trials.

9. REFERENCES

1. Breithaupt-Grogler K, Maleczyk C, Belz GG, Butzer R, Herrman V, Stass H, Wensing G. Pharmacodynamic and pharmacokinetic properties of an angiotensin II receptor antagonist – characterization by use of Schild regression techniques in man. Int J Clin Pharmacol Ther 1997; 35: 434-41.
2. Debord J, Carpentier N, Sabot C, Bertin P, Marquet P, Treves R, Merle I, Lachatre G. Influence of biological variables upon pharmacokinetic parameters of intramuscular methotrexate in rheumatoid arthritis. Int J Clin Pharmacol Ther 1998; 36: 227-30.
3. Kato Z, Fukutomi O, Yamazaki M, Kondo N, Imaeda N, Orii T. Prediction of steady-state serum theophylline concentration in children by first-order and zero -order absorption models. Int J Clin Pharmacol Ther 1994; 32: 231-4.
4. Mahmood I, Mahayni H. A limited sampling approach in bioequivalence studies: application to low half-life drugs and replicate design studies. Int J Clin Pharmacol Ther 1999; 37: 275-81.
5. Sabot C, Debord J, Roullet B, Marquet P, Merle L, Lachatre G. Comparison of 2- and 3- compartment models for the Bayesian estimation of methotrexate pharmacokinetics. Int J Clin Pharmacol Ther 1995; 33: 164-9.
6. Ulrich S, Baumann B, Wolf R, Lehmann D, Peters B, Bogerts B. Therapeutic drug monitoring of clozapine and relapse – a retrospective study of routine clinical data. Int J Clin Pharmacol Ther 2003; 41: 3-13.
7. Vrecer M, Turk S, Drinovec J, Mrhar A. Use of statins in primary and secondary prevention of coronary heart disease and ischemic stroke. Meta-analysis of randomized trials. Int J Clin Pharmacol Ther 2003; 41: 567-77.
8. Jukema JW, Bruschke AV, Van Boven AJ, Reiber JH, Bal ET, Zwinderman AH, Jansen H, Boerma GJ, Van Rappard FM, Lie KI. Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elevated serum cholesterol levels (REGRESS Trial). Circulation 1995; 91: 2528-40.
9. Antman EM, Cohen M, Bernink P, McGabe CH, Horacek T, Papuches G, Mautner B, Corbalan R, Radley D, Braunwald E. The TIMI Risk score for unstable angina pectors, a method for prognostication and therapeutic decision making. J Am Med Assoc 2000; 284: 835-42.
10. Hoetner G. The use of logit and probit models in strategic management research. Strat Mgmt J 2007; 28: 331-43.
11. Rudner LM. Computer adaptive testing. http://edres.org/scripts/cat/catdemo.htm
12. Boeckman AJ, Sheiner LB, Beal SL. NONMEM User's Guide. NONMEM Project Group, University California, San Francisco, 1992.

13. Davidian M, Giltinan DM. Nonlinear models for repeated measurements data. Chapman and Hall, NY, 1995.
14. Lindstrom MJ, Bates BM. Nonlinear mixed effects models for repeated measures data. Biometrics 1990; 46: 673-8.
15. Keusch P. Chemical kinetics, rate laws, Arrhenius equation-experiments. http://www.uniregensburg.de/Fakultaeten/nat_Fak_IV/Organische_Chemie/Didaktik.html.

CHAPTER 17

REGRESSION MODELING FOR IMPROVED PRECISION

1. INTRODUCTION

Small precision of clinical trials is defined as a large spread in the data. Repeated observations have a central tendency, but also a tendency to depart from the central tendency. If the latter is large compared to the former, the data are imprecise. This means that p-values are large, and reliable predictions cannot be made. Often a Gaussian pattern is in the data. The central tendency can, then, be adequately described using mean values as point estimates. However, if the data can be fitted to a different pattern like a linear or a curvilinear pattern, the central tendency can also be described using the best fit lines or curves of the data instead of mean values. This method is called data modeling, and may under the right circumstances reduce the spread in the data and improve the precision of the trial. Extensive research on the impact of data modeling on the analysis of pharmacodynamic / pharmacokinetic data has been presented over the past 10 years. The underlying mechanism for improved precision was explained by the late Lewis Sheiner: "Modeling turns noise into signals".[1,2] In fact, instead of treating variability as an "error noise", modeling uses the variability in the data as a signal explaining outcome. If regression models are used for such purpose, an additional advantage is the relative ease with which covariates can be included in the analysis. So far, data modeling has not been emphasized in the analysis of prospective randomized clinical trials, and special statistical techniques need to be applied including the transformation of parallel-group data into regression data. In the current chapter we demonstrate two regression models that can be used for such purpose. Both real and hypothesized examples are given.

2. REGRESSION MODELING FOR IMPROVED PRECISION OF CLINICAL TRIALS, THE UNDERLYING MECHANISM

The better the model fits the data, the better precision is obtained. Regression modeling is, essentially, an attempt to fit experimental data to specific patterns, and, subsequently, to test how far distant the data are from the best fit pattern. A statistically significant correlation indicates that the data are closer to the best fit pattern than could happen by random sampling. As an example, the simple linear regression analysis of a parallel-group study of the effects on LDL-cholesterol on pravastatin versus placebo in 884 patients, also used in the chapters 12 and 14, is

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 221–227.

given.[3] The overall spread in the data is estimated by a standard error of 0.11 mmol/l around the regression line (Figure 1 upper graph). A smaller standard error (0.024 mmol/l), and, thus, less spread in the data is provided by a multiple regression model, using two regression lines instead of one (Figure 1, lower graph). Obviously, this multiple regression pattern provided an overall shorter distance to the data than did the simple linear regression pattern. Or, in other words, it better fitted the data than did the simple linear regression. In the next few sections we give additional examples.

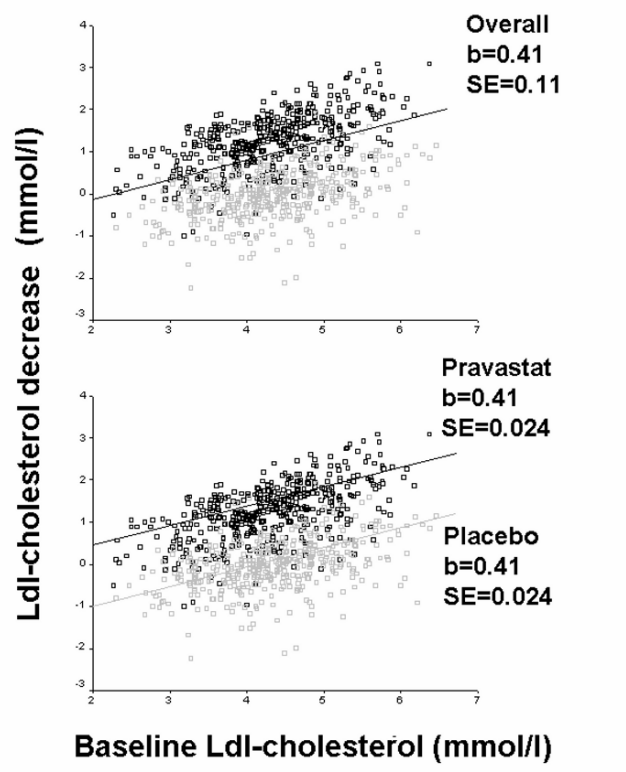


Figure 1. Linear regression analysis of parallel-group study of effect on LDL-cholesterol of pravastatin versus placebo in 872 patients. The overall spread in the data is estimated by a standard error of 0.11 mmol/l around the regression line (upper graph). The multiple regression model using two regression lines, instead of one, leads to a standard error of only 0.024 mmol/l (lower graph).

3. REGRESSION MODEL FOR PARALLEL-GROUP TRIALS WITH CONTINUOUS EFFICACY DATA

Table1 shows the data of a parallel-group trial comparing efficacy of a new laxative versus control laxative. The mean difference in response between new treatment

Table 1. a parallel-group trial comparing a new laxative versus control

patient no.	treatment modality new=0 control=1	response = stool frequency after treatment (4 week stools)	baseline stool frequency (4 week stools)
1	0	24	8
2	0	30	13
3	0	25	15
4	1	35	10
5	1	39	9
6	0	30	10
7	0	27	8
8	0	14	5
9	1	39	13
10	1	42	15
11	1	41	11
12	1	38	11
13	1	39	12
14	1	37	10
15	1	47	18
16	0	30	13
17	1	36	12
18	0	12	4
19	0	26	10
20	1	20	8
21	0	43	16
22	0	31	15
23	1	40	14
24	0	31	7
25	1	36	12
26	0	21	6
27	0	44	19
28	1	11	5
29	0	27	8
30	0	24	9
31	1	40	15
32	1	32	7
33	0	10	6
34	1	37	14
35	0	19	7

and control = 9.824 stools per 4 weeks (Se = 2.965). The t-test produces a t-value of 9.824 / 2.965 = 3.313, and the t-table gives a p-value of <0.01.

A linear regression according to
$y = a + bx$
with y = response and x = treatment modalities (0 = new treatment, 1 = control),
a = intercept, and b= regression coefficient,

produces a similar result
$b = 9.824$
$se_b = 2.965$
$t = 3.313$
p-value <0.01.

Improved precision of this data analysis is a possibility if we extend the regression model by including a second x-variable = baseline stool frequency according to

$y = a + b_1 x_1 + b_2 x_2$
with x_1 = treatment modalities (0 = new treatment, 1 = control),
x_2 = baseline stool frequencies, and b-values are partial regression coefficients.

This produces the following results
$b_1 = 6.852$
$se_{b1} = 1.792$
$t = 3.823$
p-value <0.001.

After adjustment for the baseline stool frequencies an improved precision to test the efficacy of treatment is obtained as demonstrated by a larger t-value and a smaller p-value.

4. REGRESSION MODEL FOR PARALLEL-GROUP TRIALS WITH PROPORTIONS OR ODDS AS EFFICACY DATA

Consider the underneath two by two contingency table.

	Numbers Responders	numbers non-responders
Treatment 1	30 a	45 b
Treatment 2	45 c	30 d

The odds-ratio-of-responding equals $\frac{a/b}{c/d} = \frac{30/45}{45/30} = 0.444$. The natural logarithmic (ln) transformation of this odds ratio equal -0.8110. The standard error of this logarithmic transformation is given by $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{30}+\frac{1}{45}+\frac{1}{30}+\frac{1}{45}} = 0.333$.

A t-test of these data produces a t-value of 0.8110/0.333= 2.435. According to the t-table this odds-ratio is significantly different from an odds ratio of 1.0 with a p-value of 0.015.

Logistic regression according to the model

ln odds-of-responding = a + bx
with x = treatment modality (0 or 1),
a = intercept, and b = regression coefficient,

produces the same result:

b = 0.8110
Se_b = 0.333
t-value = 2.435
p-value = 0.015

The patients can be divided into two age classes:

	Over 50 years		Under 50 years	
	responders	non-responders	responders	non-responders
Treatment 1	16	22	9	28
Treatment 2	34	4	16	21

Improved precision of the statistical analysis is a possibility if we control for age groups using the underneath multiple logistic regression model

ln odds-of-responding = $a + b_1 x_1 + b_2 x_2$
with x_1 = treatment modalities (0= treatment 1, 1= treatment 2)
x_2 = age classes (1= < 50 years, 2 = > 50 years)
b-values are regression coefficients.

The following results are obtained:

b_1 = 1.044
Se_{b1} = 0.387
t-value = 2.697
p-value = 0.007

After adjustment for age class improved precision to test the efficacy of treatment is obtained as demonstrated by a larger t-value and smaller p-value.

5. DISCUSSION

Multiple regression analysis of confounding variables, although routinely used in retrospective observational studies, is not emphasized in prospective randomized clinical trials (RCTs). The randomization process ensures that such potential

confounders are equally distributed among the treatment groups. If not, the result of the study is flawed, and regression analysis is sometimes used in a post hoc attempt to salvage the data, but there is always an air of uncertainty about such data. Multiple regression can, however, be used in prospective RCTs for a different purpose. Certain patient characteristics in RCTs may cause substantial spread in the data even if they are equally distributed, and, thus, independent of the treatment groups. Including such data in the efficacy analysis may reduce the overall spread in the data, and reduce the level of uncertainty in the data analysis. Regression models are also adequate for such purpose, although rarely applied so far.

Regression modeling is a very sophisticated statistical technique which needs to be applied carefully and under the right circumstances. Therefore, when using regression analysis for the purpose of improving precision of RCTs a number of potential problems have to be accounted. They have been recently published by us[4], are reviewed in the previous chapter.

1. The sensitivity of testing is jeopardized if the linear or exponential models do not fit the data well enough. This can be checked for example by scatter-plots and histograms.
2. Relevant clinical phenomena like unexpected toxicity effects and complete remissions can go unobserved by the use of a regression model to assess the data.
3. The inclusion of multiple variables in regression models raises the risk of clinically unrealistic results.

Nonetheless, if certain patient characteristics are largely independent of the treatment modality, they can be included in the data analysis, in order to reduce the overall spread in the data. We should emphasize that it has to be decided prior to the trial and stated explicitly in the trial protocol whether a regression model will be applied, because post hoc decisions regarding regression modeling like any other post hoc change in the protocol raises the risk of statistical bias due to multiple testing. Naturally, there is less opportunity for modeling in a small trial than in a large trial. There is no general rule about which sample sizes are required for sensible regression modeling, but one rule-of-thumb is that at least ten times as many patients are required as the number of variables in the model. This would mean that a data set of at least 30 is required if we wish to include a single covariate in the model for the purpose of improving precision. With every additional covariate in the model an extra regression weight must be estimated, which may lead to a decreased rather than improved precision. Regression analysis can be adequately used for improving precision of efficacy analysis. Application of these models is very easy since many computer programs are available. For a successful application the fit of the regression models should, however, always be checked, and the covariate selection should be sparse.

6. CONCLUSIONS

Small precision of clinical trials is defined as a large spread in the data. Certain patient characteristics of randomized controlled trials may cause substantial spread in the data even if they are equally distributed among the treatment groups. The objective of this chapter was to assess whether improved precision of the analysis can be obtained by transforming the parallel-group data into regression data, and, subsequently, including patient characteristics in the analysis.

In a 35 patient parallel-group trial with continuous efficacy data, after adjustment of the efficacy scores for baseline scores, the test-statistic rose from $t = 3.313$ to $t = 3.823$, while the p-value fell from < 0.01 to < 0.001. In a 150 patient parallel-group trial with odds as efficacy variable, after adjustment of the efficacy variable for age class, the test statistic rose from $t = 2.435$ to $t = 2.697$, while the p-value fell from 0.015 to 0.007.

We conclude that regression analysis can be adequately applied for improving precision of efficacy data of parallel-group trials. We caution that, although application of these models is very easy with computer programs widely available, the fit of the regression models should always be carefully checked, and the covariate selection should be sparse.

7. REFERENCES

1. Sheiner LB, Steimer JL. Pharmacokinetic / pharmacodynamic modeling and drug development. Clin Pharmacol Ther 1984; 35: 733-41.
2. Fuseau E, Sheiner LB. Simultaneous modeling of pharmacokinetics and pharmacodynamics with a nonparametric pharmacodynamic model. Clin Pharmacol Ther 1984; 35: 733-41.
3. Cleophas TJ. The sense and non-sense of regression modeling for increasing precision of clinical trials. Clin Pharmacol Ther 2003; 74: 295-7.
4. Cleophas TJ. Problems in regression modeling of randomized clinical trials. Int J Clin Pharmacol Ther 2005; 43: 5-12.

CHAPTER 18

POST-HOC ANALYSES IN CLINICAL TRIALS, A CASE FOR LOGISTIC REGRESSION ANALYSIS

1. MULTIPLE VARIABLES METHODS

Multiple varaibles methods are used to adjust asymmetries in the patient characteristics in a trial (see page 171 for a discussion of the difference between multivariate and multiple variables methods). It can also be used for a subsequent purpose. In many trials simple primary hypotheses in terms of efficacy and safety expectations, are tested through their respective outcome variables as described in the protocol. However, sometimes it is decided already at the design stage that post hoc analyses will be performed for the purpose of testing secondary hypotheses. E.g., suppose we first want to know whether a novel beta-blocker is better than a standard beta-blocker, and second, if so, whether this better effect is due to a vasodilatory property of the novel compound. The first hypothesis is assessed in the primary analysis. For the second hypothesis, we can simply adjust the two treatment groups for difference in vasodilation by multiple regression analysis and see whether differences in treatment effects otherwise are affected by this procedure. However, with small data power is lost by such procedure. More power is provided by the following approach. We could assign all of the patients to two new groups: patients who actually have improvement in the primary outcome variable and those who have not, irrespective of the type of beta-blocker. We, then, can perform a regression analysis of the two new groups trying to find independent determinants of this improvement. If the dependent determinant is binary, which is generally so, our choice of test is logistic regression analysis. Testing the second hypothesis is, of course, of lower validity than testing the first one, because it is post-hoc and makes use of a regression analysis which does not differentiate between causal relationships and relationships due to an unknown common factor.

2. EXAMPLES

In a double-blind randomized study of the new beta-blocker celiprolol for patients with angina pectoris the main outcome variable was anginal attack rate. Additional outcome variables include systolic and diastolic blood pressure, heart rate, rate pressure product, peripheral vascular resistance. Although this study measures several outcomes, the various outcomes to some degree measure the same thing, and this may be particularly so with blood pressure, heart rate and pressure rate product since they are assumed to represent oxygen demand to the heart, which is

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 229–234.

jeopardized during anginal attacks. The new beta-blocker has been demonstrated preclinically not only to reduce rate pressure product like any other beta-blocker but also to reduce peripheral vascular resistance. The novel beta-blocker indeed performed significantly better than the latter (persistent angina pectoris at the completion of the trial 17 versus 33 %, $P<0.01$, $1-\beta = \pm 80\%$), and this was accompanied by a significantly better reduction of systolic blood pressure and reduction of peripheral resistance. A problem with multiple variables analysis is its relatively small power with usual sample sizes. For the purpose of better power patients may be divided into new groups according to their main outcome. In order to determine the most important determinants of the better clinical benefit, the patients were, therefore, divided into two new groups: they were assigned to "no-angina-pectoris" at the completion of the trial or "persistent-angina-pectoris" (table 1). The univariable analysis of these two new groups showed that most of the

Table 1. Angina pectoris and odds ratios of persistent angina pectoris in the celiprolol (novel compound) and propranolol (reference compound) group adjusted for independent variables (data from Cleophas et al; Clin Pharmacol Ther 1996; 45: 476). Odds ratio = odds of persistent angina pectoris in the celiprolol group / odds of persistent angina pectoris in the propranolol group. Means ± SDs are given

	No angina pectoris (n=23) mean ± SD	P	persistent angina pectoris (n=30) mean ± SD
systolic blood pressure (mm Hg)	134 ± 17	<0.001	155 ± 19
diastolic blood pressure (mm Hg)	77 ± 13	<0.02	84 ± 9
heart rate (beat/min)	65 ± 9	<0.09	69 ± 9
rate pressure product (mm Hg.Beats/min.10^{-3})	8.6 ± 11	<0.001	10.7 ± 14
fore arm blood flow (ml/100ml tissue.min)	8.8 ± 10.8	<0.02	4.1 ± 2.2
treatment assignment (celiprolol / propanolol)	18 / 5	<0.001	8 / 22

odds ratio of persistent angina		95% CIs	P-value
unadjusted	0.38	0.25 - 0.52	< 0.002
adjusted for rate pressure product	0.13	0.05 - 0.22	< 0.0005
adjusted for systolic pressure plus heart rate	0.12	0.04 - 0.20	< 0.0005

CI = confidence interval; SD = standard deviation.

additional outcome variables including treatment assignment were significantly different between the two groups. These variables were entered in the logistic

regression analysis: the variables double product, systolic blood pressure and heart rate were independent of treatment assignment, while fore arm blood flow (=1/peripheral vascular resistance) was not. After adjustment for fore arm blood flow the difference in treatment assignment was lost. This suggests that celiprolol exerted its beneficial effect to a large extent through its peripheral vasodilatory property.

As a second example is given a double-blind randomized parallel-group study comparing chronotropic (mibefradil ands diltiazem) and non-chronotropic calcium channel blockers (amlodipine) in patients with angina pectoris. Although all of the calcium channel blockers improved exercise tolerance as estimated by % increased time to onset ischemia during bicycle ergometry, mibefradil and diltiazem performed better than amlodipine (20.8 and 12.4 s versus 9.9 s, $P < 0.01$ and < 0.001). In order to determine the most important determinants of this better clinical benefit, patients were divided into two new groups: they were assigned to non-responders if their change in ischemic onset time was zero or less, and to responders if it was larger than zero (table 2). Univariable analysis of these two groups showed that many variables including treatment assignment were significantly different between the two groups.

Table 2. Mean data (SDs) after assignment of patients according to whether (responders) or not (non-responders) their ischemia-onset-time increased after treatment with calcium channel blockers, and odds ratios of mibefradil or diltiazem versus amlodipine for responding, unadjusted and after adjustment for difference of heart rate (Cleophas et al; Br J Clin Pharmacol 1999; 50: 545). Odds ratio = odds of responding on mibefradil or diltiazem or amlodipine / odds of responding on amlodipine

	responders (n=239) mean (SD)	non-responders (n=61) mean (SD)	P-value
at rest			
systolic blood pressure (mm Hg)	-5 (19)	-1 (23)	0.27
diastolic blood pressure (mm Hg)	-5 (10)	-3 (10)	0.13
heart rate (beats/min)	-5 (11.0)	1.1 (9.6)	< 0.001
rate pressure product (mm Hg.beats/min.10^{-3})	-1.0 (1.9)	0.1 (2.1)	< 0.001
at maximal workload			
systolic blood pressure (mm Hg)	-1 (21)	-2 (27)	0.68
diastolic blood pressure (mm Hg)	-4 (11)	-4 (11)	0.97
heart rate (beats/min)	-12 (17)	-6 (15)	0.010
rate pressure product (mm Hg.beats/min.10^{-3})	-2.3 (4.5)	-1.2 (4.5)	0.090
treatment assignment (n, %)			
amlodipine	76 (32%)	27 (44%)	
diltiazem	75 (31%)	26 (43%)	
mibefradil	88 (37%)	8 (13%)	

	unadjusted odds ratio (95% CIs)	odds ratio adjusted for change in heart rate (95% CIs)
amlodipine	1 (-)	1 (-)
diltiazem	1.02 (0.55-1.92)	0.86 (0.45-1.66)
mibefradil	3.91 (1.68-9.11)	2.26 (0.86-5.97)

CI = confidence interval; SD = standard deviation.

These variables were entered into the logistic regression analysis: the difference in treatment assignment between the two groups was lost after adjustment for heart rates. This suggests that the beneficial effect of calcium channel blockers in this category of patients is largely dependent upon their effect on heart rate.
It is important to recognize that in the first study there is a positive correlation between peripheral flow and clinical benefit (when peripheral flow increases benefit gets better), whereas in the second study there is a negative correlation between heart rate and clinical benefit (when heart increases benefit gets worse). Multiple variables analysis only measures dependencies but makes no differences between a positive and negative correlation. So, we must not forget to look at the trend in the data before interpretations can be made.

3. LOGISTIC REGRESSION EQUATION

Logistic regression is similar to linear regression the main equation of which is explained in chapter 11:

$$y = a+b_1x_1+b_2x_2+\ldots b_nx_n$$

Linear regression software finds for you an equation that best predicts the outcome y from one or more x variables. Continuous data are measured. Y is assumed to be the expected value of a normal distribution. With y being a binary (yes/no) variable, the proportion of, e.g., "yes" data (p in the underneath example) lies between 0 and 1, and this is too small a range of values for the expression of a summary of multiple variables like $a+b_1x_1+b_2x_2+\ldots b_nx_n$. The range of y-responses can be broadened to 0 to ∞ if we take p/(1-p) as y-variable , and even to $-\infty$ to $+\infty$ if we take ln p/(1-p). The simplest logistic regression model using only a single x-variable can be presented in a contingency table of proportional data:

	high.....	low leucocyte count
Transplant rejections	p_1	$1-p_1$
No transplant rejections	p_0	$1-p_0$

$$\ln \frac{p_1}{1 - p_1} = bx + a$$

$$\text{If } x = 1 \rightarrow \ln \frac{p_1}{1 - p_1} = b + a$$

$$\text{If } x = 0 \rightarrow \ln \frac{p_0}{1 - p_0} = a$$

$$\ln \frac{\frac{p_1}{1 - p_1}}{\frac{p_0}{1 - p_0}} = b$$

Odds ratio = e^b

p/(1-p) = the odds of finding yes-data. The regression coefficient value (b-value) in the logistic regression equation can be best understood as the natural logarithm of the odds ratio of finding $p_1/(1-p_1)$ given $p_0/(1-p_0)$. Although with multiple-variables logistic regression becomes a formidable technique, it is straightforward to understand, and logistic regression increasingly finds its way into the secondary analysis of trial data.

4. CONCLUSIONS

Sometimes it is decided already at the design stage of a clinical trial to perform post-hoc analyses in order to test secondary hypotheses. For the purpose of power we may make two new groups: those who have improvement and those who have not, irrespective of the type of treatment. We, then, can perform a regression analysis of the two new groups trying to find independent determinants of improvement. If one or more determinants for adjustment are binary, which is generally so, our choice of test is logistic regression analysis. This procedure does of course provide no proof. However, it may give strong support for the presence of particular underlying mechanisms in the data.

5. REFERENCES

1. Cleophas TJ, Remmert HP, Kauw FH. Celiprolol versus propranolol in unstable angina pectoris. Clin Pharmacol Ther 1996; 45: 476-3.
2. Van der Vring AF, Cleophas TJ, Zwinderman AH, et al. Different classes of calcium channel blockers in addition beta-blockers for exercise induced angina pectoris. Br J Clin Pharmacol 1999; 50: 545-60.

CHAPTER 19

CONFOUNDING

1. INTRODUCTION

When published, a randomized parallel-group drug trial essentially includes a table listing all of the factors, otherwise called baseline characteristics, known possibly to influence outcome. E.g., in case of heart disease these will probably include apart from age and gender, the prevalence in each group of diabetes, hypertension, cholesterol levels, smoking history, other cardiovascular comorbidities, and concomitant medications. If the prevalence of such factors is similar in the two groups, then we can attribute any difference in outcome to the effect of test-treatment over reference-treatment. However, if this is not the case, we have a problem which can be illustrated by an example. Figure 1 shows the results of a

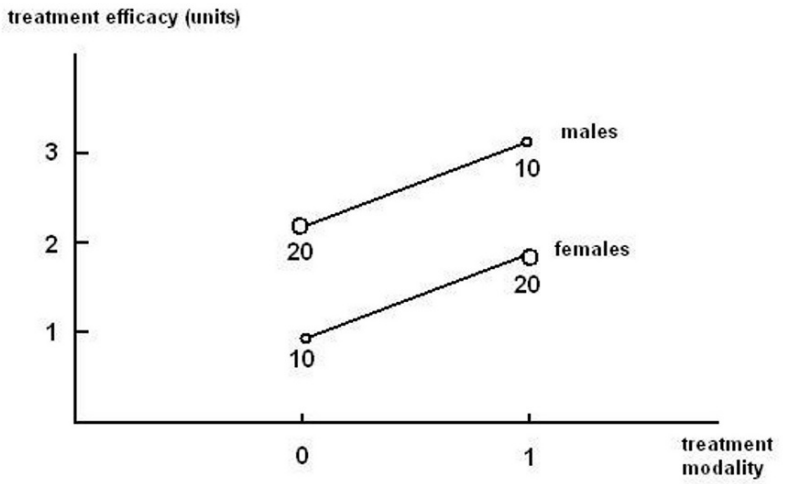


Figure 1. Efficacy of control (0) and test treatment (1) in a trial where females and males are assessed separately. The magnitude of the circles corresponds to the size of the subclass samples.

study where the treatment effects are better in the males than they are in the females. This difference in efficacy does not influence the overall assessment as long as the numbers of males and females in the treatment comparison are equally distributed. If, however, many females received the new treatment, and many males received the control treatment, a peculiar effect on the overall data analysis is observed: the overall regression line is close to horizontal, giving rise to the erroneous conclusion that no difference in efficacy exists between treatment and

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 235–244.

control. This phenomenon is called confounding, and may have a profound effect on the outcome of a trial. In randomized controlled trials confounding is, traditionally, considered to play a minor role in the data. The randomization ensures that no covariate of the efficacy variable is associated with the randomized treatment.[1] However, the randomization may fail for one or more variables, making such variables confounders. Then, adjustment of the efficacy estimate should be attempted. Methods include subclassification[2], regression modeling[1], and propensity scores.[3,4] This paper reviews these three methods and uses hypothesized and real data examples for that purpose.

2. FIRST METHOD FOR ADJUSTMENT OF CONFOUNDERS: SUBCLASSIFICATION ON ONE CONFOUNDER

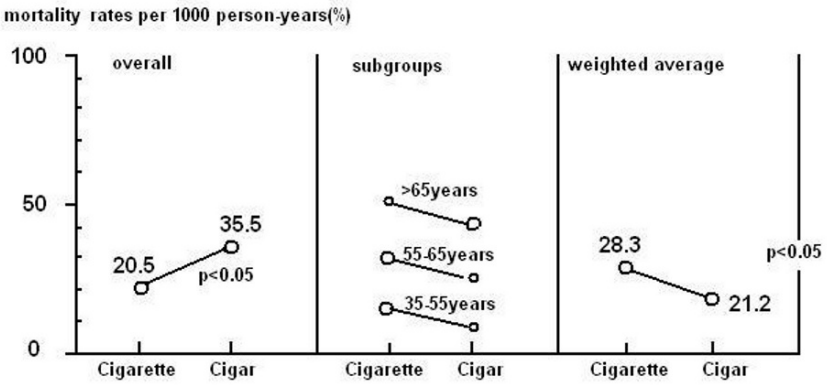


Figure 2. Example of subclassification on one confounder (age). Left graph: overall mortality from cigar smoking is significantly larger than the from cigarette smoking; middle graph: if divided into three subclasses, the mortality from cigarettes is larger than from cigars although the differences were not statistically significant; right graph: a weighted average from the comparisons from the middle graph shows that the mortality from cigarettes is significantly larger than from cigars.

Figure 2 gives an example of confounding on one variable, age. In a large database from Canada[2] the overall mortality from cigar smoking was significantly larger than that from cigarette smoking However, the cigar smokers were significantly older, and, therefore, hardly comparable (mean age 66 versus 51 years, $p<0.05$). How do we assess this inequality of age in the groups. One way of assessment is as follows: (1) we divide the population into age subclasses of approximately equal size, the younger, middle-ages, and older, then, (2) compare mortality per subclass, and, finally, (3) calculate a so-called weighted average. Figure 2 shows that in any

of the three subclasses mortality from cigarettes was higher than from cigars, but differences were not statistically significant. The higher mortality from cigars in the overall assessment was caused by the fact that many youngsters smoked cigarettes, while many elderly, obviously, preferred cigars. The weighted average is calculated as

$$\frac{(R_{cigt} - R_{cig})_1 / \text{variance}_1 \; + \; (R_{cigt} - R_{cig})_2 / \text{variance}_2 \; + \; (R_{cigt} - R_{cig})_3 / \text{variance}_3}{1 / \text{variance}_1 \; + \; 1 / \text{variance}_2 \; + \; 1 / \text{variance}_3}$$

where $(R_{cigt} - R_{cig})$ = difference in mortality rate (R) between cigarette and cigar smokers for subclass 1 (the younger), subclass 2 (the middle-ages), and subclass 3 (the elderly) respectively, variance are the variances of these difference-in-rates. For testing the significance of difference between cigarette and cigar smoking of the weighted averages a weighted variance is required which is calculated as add-up sum of the separate variances. Cochran used, among other examples, the above example and reasoned that as long as a reasonable number of persons are in each subclass this procedure removes up to 90% of the bias due to confounding.[2] The advantages of subclassification over regression analysis for confounding include, first, that empty subclasses in the treatment comparisons are readily vizualized, and, second, that subclassification does not rely on a linear or other regression model, and is, thus, universally applicable. The problem with subclassification is that, with multiple confounders, it is simply impossible to divide the population in subclasses. For that purpose multivariable regression analysis is required.

3. SECOND METHOD FOR ADJUSTMENT OF CONFOUNDERS: REGRESSION MODELING

Instead of subclassification regression modeling can be applied to adjust a confounding variable.[1] An example is given in Figure 1. The data of a parallel-group study produced a significant difference

Mean treatment 0	1.666	standard deviation	0.479
Mean treatment 1	2.333	standard deviation	0.479
Difference	0.666	standard error 0.214	$p<0.001$

The same result is obtained using a linear regression model with treatment modality on the x-axis and treatment efficacy on the y-axis. The regression coefficient is the direction coefficient of the regression line and equals 0.666 (standard error 0.214), which is equal to the mean treatment efficacy as obtained in the above usual analysis. From the Figure 1 it is concluded that gender is a confounding variable and the data are thus adjusted for gender by adding it as a second dependent variable (variable z) to the model. SPSS[5] statistical software produces the following results after commanding: statistics; regression; linear;

	r^2	b	se	P-value
Unadjusted	0.333	0.666	0.214	<0.001
Adjusted	1.000	1.000	0.000	<0.00001

where r = correlation coefficient ; b = regression coefficient; se = standard error.

The adjusted efficacy estimate b may become smaller or larger than the unadjusted estimate, depending on the direction of the associations of the confounder with the randomized treatment and the efficacy variable. Let b_1 and b_1^* denote the unadjusted and the adjusted efficacy estimate, and let r_{xz} and r_{yz} be the correlations of the confounder (z) with the randomized treatment (x) and the efficacy variable (y), then the following will hold:

if	$r_{xz}>0$ and $r_{yz}>0$	then	$\lvert b_1^* \rvert < \lvert b_1 \rvert$,
if	$r_{xz}>0$ and $r_{yz}<0$	then	$\lvert b_1^* \rvert > \lvert b_1 \rvert$,
if	$r_{xz}<0$ and $r_{yz}<0$	then	$\lvert b_1^* \rvert < \lvert b_1 \rvert$,
if	$r_{xz}<0$ and $r_{yz}>0$	then	$\lvert b_1^* \rvert > \lvert b_1 \rvert$,

Notice the possibility that the unadjusted efficacy estimate b_1 is zero whereas the adjusted estimate b_1^* is unequal to zero: an efficacy-difference between treatments may be masked by confounding. In clinical trials it is sensible to check the balance between treatment groups of all known covariates of the efficacy variable. In most trials there are many more covariates and one should be careful to consider as a confounder a covariate which was not reported in the literature before. The advantage of regression analysis compared to subclassification is that multiple variables can be added to the model in order to test whether they are independent determinants, and thus significant confounders of the dependent variable, treatment efficacy. The power of these tests is a sensitive function of the number of patients in the trial. Naturally, there is less opportunity for modeling in a small trial than there is in a large trial. There is no general rule about what sample sizes are required for sensible regression modeling, but a rule of thumb is that at least ten times as many patients are required as the number of covariates in the model. If these requirements are not met, the trial rapidly loses power, and a different approach is needed. Propensity scores have been recommended for that purpose.

4. THIRD METHOD FOR ADJUSTMENT OF CONFOUNDERS: PROPENSITY SCORES

The method of propensity scores is relatively new (1983, Rosenbaum and Rubin)[3,4], but increasingly accepted in observational research, although its theoretical properties have not yet been entirely elucidated. Each patient is

assigned a propensity score, which is his/her probability, based on his/her covariate value, of receiving a particular treatment modality. As an example, in a parallel group study of 100 versus 100 patients, 63 out of 100 patients in treatment group 1 were older than 65, while 76 were so in treatment group 2. The probability of receiving treatment 1 in patients older than 65 years can be calculated to be 63/76 / 37/24 = 0.54. This probability equals the odds of treatment 1 with the characteristic / odds of treatment 1 without the characteristic, otherwise called the odds ratio (OR) of the two. This odds ratio can, then, be applied as measure for adjustment the asymmetric prevalence of the patient characteristic between the treatment groups. Two alternative methods as described in the above sections are available, and propensity score are therefore rarely used for that purpose. Things are different when multiple confounding variables are in a treatment comparison. Subclassification is, then, impossible, and regression modeling gets powerless.
Propensity scores including more than 1 covariates can be calculated according to the following method (Table 1). For each patient the odds ratios of the covariates at

Table 1. With propensity scores for each patient the odds ratios of the covariates at risk of confounding are calculated (odds ratio = odds of treatment 1 with confounder / odds without confounder) (upper table). Then the statistically significant odds ratios are assumed to be significant confounders and are combined into one propensity score per patient calculated as their product of multiplication (lower table).

Characteristic at risk of confounding	treatment 1 n =100	treatment 2 n =100	Odds treatment 1 with characteristic / odds without	p-value
1. Age>65 years	63	76	0.54 (63/76 / 37/24)	0.05
2. Age<65 years	37	24	1.85 (= 1/ $OR_{Age>65\ years}$)	0.05
3. Diabetes	20	33	0.51	0.10
4. No diabetes	80	67	1.96	0.10
5. Smoker	50	80	0.25	0.10
6. No smoker	50	20	4.00	0.10
7. Hypertension	60	65	0.81	ns
8. No hypertension	40	35	1.23	ns
9. Cholesterol	75	78	0.85	ns
10.No cholesterol	25	22	1.18	ns
11.Renal insufficiency	12	14	0.84	ns
12.No renal insufficiency	88	86	1.31	ns

	old y/n	dm y/n	smoker y/n	propensity score = OR_1 xOR_2 xOR_3
Patient 1	y	y	n	0.54 x 0.51 x 4 = 1.10
2	n	n	n	1.85 x 1.96 x 4 = 14.5
3	y	n	n	0.54 x 1.96 x 4 = 3.14
4	y	y	y	0.54 x 0.51 x .025 = 0.06885
5	n	n	y	
6	y	y	y	
7	….			

OR = odds ratio; y = yes ; n = no; ns = not significant; p<0.01 = statistically significant here.

risk of confounding are calculated. Statistically significant odds ratios are assumed to be significant confounders (Table 1 upper table), and are, subsequently, combined into one propensity per patient in the form of their product of multiplication (Table 1 lower table). The next step is to divide the patients into four or more subclasses dependent on their magnitude of propensity score. Then, calculate per subclass mean difference in treatment effect. In order to determine an adjusted overall treatment difference between the two treatment groups, a weighted average can be calculated using the same weighting procedure as that used with subclassification described in one of the previous sections. Figure 3 gives the

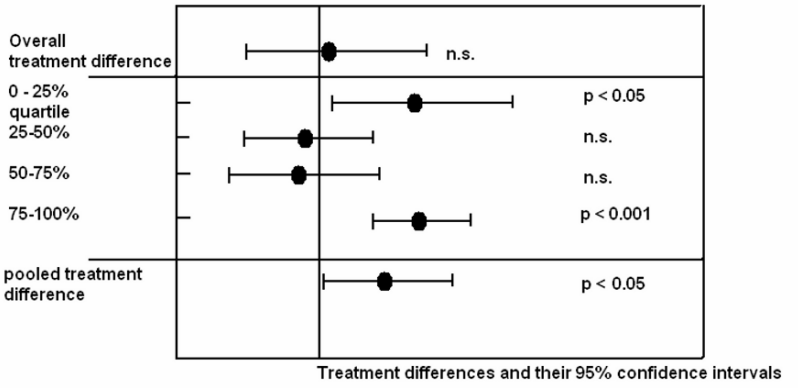


Figure 3. A propensity score adjustment for confounding. The patients are divided into 4 quartiles according to the magnitude of their propensity scores. The weighted average (pooled treatment difference) is statistically significant different from zero, while the overall treatment difference (unadjusted for confounders) was not so.

results of this procedure. It is observed that the adjusted overall difference is larger than the unadjusted overall difference, and unlike the latter the former is statistically significantly different from a difference of zero. Obviously, the confounders masked a true treatment difference, which is being unmasked by the propensity procedure. As an alternative to the subclassification procedure, a regression model comparable with the regression, described in the above section, with treatment efficacy as independent, treatment modality as first dependent and propensity score as second dependent variable will produce a largely similar result.

5. DISCUSSION

In large randomized controlled trials a random imbalance of the covariates is mostly negligible. However, with smaller studies it may be substantial. In the latter situation assessment and adjustment for confounders is a requirement in order to reduce a biased assessment of the treatment comparison. In the current paper three methods for that purpose are reviewed.

We like to discuss the limitations of the three methods for the assessment of confounding. The subclassification assessment has the limitation that it can only be used for one confounder. With multiple confounders multivariable regression analysis is the method of choice. However, this method is limited by the sample size of the trial. We need at least ten times as many patients in our samples than numbers of variables in the analysis. This would mean that, with n = 100 in either of two subgroups, and the treatment efficacy and treatment modality as primary variables, we have only room for 8 additional variables. For studies with binary efficacy variables or survival studies other regression models are adequate like multivariable logistic or multivariable Cox regression. However, such models often require additional primary variables like a variable for censored data, and even less room is left for additional variables for the purpose of confounding assessments. With multiple covariates at risk of confounding, propensity scores is an alternative possibility.[3,4] However, also the method of propensity scores has major limitations. First, propensity score are entirely based upon odds ratios, and odds ratios are relative rather than absolute measures.[6] E.g., if one patient in treatment group 1 and two patients in treatment group 2 have a certain characteristic, the odds of treatment 1 with the characteristic / the odds of treatment 1 without the characteristic is 0.5, which is a huge odds ratio (OR) for an otherwise insignificant covariate. It has been advocated to include in confounding assessments any variable that is potentially causally related to the treatment response.[7] However, technically, statistically insignificant ORs in a propensity score severely reduce the power of the method, and regression models may provide better sensitivity under these circumstances as demonstrated by Soledad Cepeda et al in a Monte Carlo simulation study.[8] Second, very large and very small ORs are not reliable predictors of the chance of a patient being in a category. If such ORs are included in propensity scores a simulated atmosphere of certainty is created. Nonetheless, propensity scores that account the above limitations, and include a sensible series of covariates relevant to the treatment comparison according to

previous knowledge, can be more reliable than multivariable regression modeling for adjustment of covariates, particularly if studies are not large.

Irrespective of the method of adjustment for confounders, the question is should we adjust or not. Wickramaratne and Holford [9] gave an example in which identical results were obtained whether or not account was taken of the potential confounder. The variance estimated from the collapse table (ignoring the confounder) was lower than that from the stratified table. They concluded that precision can be lost by unnecessary adjusting for covariates. Studies at risk of this phenomenon are of course particularly those with small samples and wide variances in the subgroups. In addition, no major differences between the variances of the covariates included in the analysis is an important requirement for assessing causal effects as attempted in most clinical trials.[10]

For the assessment of confounding the intention to treat population, unlike the completed protocol population, has the advantage that samples are larger. The problem is that treatment differences may be smaller and precision may be lost. Precision may be somewhat improved using the so-called least observation carried forward analysis under the assumption that the last observation is the best estimate for the missing data.[11] There are often many candidates for inclusion as covariates, but the choice should be made a priori and specified in the protocol. If subgroups are identified post-hoc, the exploratory nature of the subgroups analyses should be emphasized and the subgroup issue should be further assessed in subsequent independent and prospective data-sets.

Sometimes in clinical trials time-concentration relationships of new drugs are assessed. These assessments make use of multi-exponential rather linear regression models. As no direct methods for the analysis of exponential models are available, data have to be transformed and Laplace's transformations are often used for that purpose.[12] The Laplace transformed relationships are linear or quadratic and can be analyzed and adjusted for confounders using linear or polynomial regression analysis. Statistical software for that purpose includes S-plus SAS[13] and the Nonmem (non-linear mixed effects models) Software.[14]

We should add that all of the methods described in this paper can not be used for the assessment or adjustment of interacting factors. Unlike confounding where all of the treatments perform better in one subclass than in the other, interaction shows that one treatment outperforms in one subclass while the other treatment does so in the other. The presence of interaction can be statistically tested by comparing the effect sizes, e.g., by using odds ratios of treatment success in either subclass, or by mixed models analysis of variance.[15]

6. CONCLUSIONS

In large randomized controlled trials the risk of random imbalance of the covariates is mostly negligible. However, with smaller studies it may be substantial. In the latter situation assessment and adjustment for confounders is a requirement in order to reduce a biased assessment of the treatment comparison. The objective of

this chapter is to review three methods for confounding assessment and adjustment for a nonmathematical readership.

First method, subclassification: the study population is divided into subclasses with the same subclass characteristic, then, treatment efficacy is assessed per subclass, and, finally, a weighted average is calculated.

Second method, regression modeling: in a multivariable regression model with treatment efficacy as independent and treatment modality as dependent variable, the covariates at risk of confounding are added as additional dependent variables to the model. An analysis adjusted for confounders is obtained by removing the covariates that are not statistically significant.

Third method, propensity scores: each patient is assigned several odds ratios (ORs), which are his/her probability, based on his/her covariate value of receiving a particular treatment modality. A propensity score per patient is calculated by multiplying all of the statistically significant ORs. These propensity scores are, then, applied for confounding adjustment using either subclassification or regression analysis.

The advantages of the first method include that empty subclasses in the treatment comparison are readily visualized, and that subclassification does not rely on a linear or any other regression model. A disadvantage is, that it can only be applied for a single confounder at a time. The advantage of the second method is, that multiple variables can be included in the model. However, the number of covariates is limited by the sample size of the trial. An advantage of the third method is, that it is generally more reliable and powerful with multiple covariates than regression modeling. However, irrelevant covariates and very large / small ORs reduce power and reliability of the assessment. The above methods can not be used for the assessment of interaction in the data.

7. REFERENCES

1. Cleophas TJ, Zwinderman AH, Cleophas AF. Statistics applied to clinical trials. Springer, New York 2006, pp 141-50.
2. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968; 24: 295-313.
3. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55.
4. Rubin DB. Estimating causal effects from large data sets using propensity score. Ann Intern Med 1997; 127: 757-63.
5. SPSS Statistical Software. http:/www.spss.com
6. Sobb M, Cleophas TJ, Hadj-Chaib A, Zwinderman AH. Clinical trials: odds ratios, why to assess them, and how to do so. Am J Ther 2008; 15: 44-53.
7. Cleophas TJ, Tuinenburg E, Van der Meulen J, Kauw FH. Wine drinking and other dietary characteristics in males under 60 before and after acute myocardial infarction. Angiology 1996; 47: 789-96.
8. Soledad Cepeda M, Boston R, Farrer JT, Strom BL. Comparison of logistic regression versus propensity scores when the number of events is low and there are multiple confounders. Am J Epidemiol 2003; 158: 280-7.

9. Wickramaratne PJ, Holford TR. Confounding in epidemiological studies: the adequacy of the control groups as a measure of confounding. Biometrics 1987; 43: 751-65.
10. Huppler Hullsiek K, Louis TA. Propensity scores modeling strategies for the causal analysis of observational data. Biostat 2002; 3: 179-93.
11. Begg CB. Commentary: ruminations on the intent to treat. Control Clin Trials 2000; 21: 241-3.
12. Beal SL, Sheiner LB. A note on the use of Laplace's approximations for non -linear mixed -effects models. Biometrika 1996; 83: 447-52.
13. SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas
14. Boeckman AJ, Sheiner LB, Beal SL. 1984 NONMEM user guide: part V. NONMEM Project Group, University of California, San Francisco.
15. Cleophas TJ, Zwinderman AH, Cleophas AF. Statistics applied to clinical trials. Springer, New York 2006, 329-36.

CHAPTER 20

INTERACTION

1. INTRODUCTION

In pharmaceutical research and development, multiple factors like age, gender, comorbidity, concomitant medication, genetic and environmental factors co-determine the efficacy of the new treatment. In statistical terms we say they interact with the treatment efficacy. It is impossible to estimate all of these factors. Instead, randomized controlled trials are used to ensure that no major imbalances exist regarding these factors, and an overall assessment is made. The limitation of this approach becomes obvious once the new medicine is applied in practice where benefits of new medicines are far less consistent than they are in the trials.[1] Despite this limitation, interaction effects, are not routinely assessed in clinical trials, probably because the statistical methods for identifying and integrating them into the data have low power. Moreover, if we introduce a large number of interaction terms in a regression analysis, the power to demonstrate a statistical significance for the primary endpoint will be reduced. Nonetheless, the assessment of a small number of interaction terms in clinical research can be an important part of the evaluation of new drugs, particularly, if it can be argued that the interaction terms make clinically sense. The current chapter gives some important factors that may interact with the treatment efficacy, and proposes some guidelines for implementing an interaction assessment in the analysis of clinical trials, in order to better predict the efficacy / safety of new medicines in future clinical treatment of individual patients.

2. WHAT EXACTLY IS INTERACTION, A HYPOTHESIZED EXAMPLE

The aim of clinical trials of new medicines is, generally, to use the estimated effects in forecasting the results of applying a new medicine to the general population. For that purpose a representative sample of subjects is treated with the new medicine or a control medicine. For example, in a parallel group study 400 patients are treated as follows:

	patients who received new medicine (n=200)	control medicine (n=200)	p-value
successfully treated patients	130/200 (65%)	110/200 (55%)	<0.01.

Based on this assessment the best bet about the difference between the two treatment modalities is given by the overall difference between the two treatment groups. We can expect that the new medicine performs 10% better than does the

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 245–261.

control medicine. If, however, we include the factor gender into our data, the results look slightly different:

	patients who received new medicine (n=200)	control medicine (n=200)	accumulated data
successfully treated females	55/100	65/100	120/200
successfully treated males	75/100	45/100 +	120/200
	130/200	110/200	

The above result shows, that, although no difference between females and males exists in the accumulated data, the new medicine performs better in the males, while the control medicine does so in the females. The adequate interpretation of this result is, if you don't wish to account gender, then the new medicine performs better, while, if you include only females, the control medicine performs better. The treatment modalities interact with gender. Interaction effects usually involve situations like this. It is helpful to display interaction effects important to the interpretation of the data in a graph with treatment modality on the x-axis and subgroup results on the y-axis. If the lines drawn for each subgroup are parallel (Figure 1 upper graph), no interaction is in the data. A different slope, and, particularly, crossing lines (Figure 1, lower graph), suggest the presence of interaction effects between treatment efficacy and subgroups, in our example *treatment x gender interaction*. The new medicine is better in females than it is in males.

The medical concept of interaction is synonymous to the terms heterogeneity and synergism. Interaction must be distinguished from confounding. In a trial with interaction effects the parallel groups have similar characteristics. However, there are subsets of patients that have an unusually high or low response. With confounding things are different. For whatever reason the randomization has failed, the parallel groups have asymmetric characteristics. E.g., in a placebo-controlled trial of two parallel-groups asymmetry of age may be a confounder. The control group is significantly older than the treatment group, and this can easily explain the treatment difference. Particularly, in survival studies differences in baseline age may be an important confounder as recently demonstrated by De Craen and Westendorp.[2]

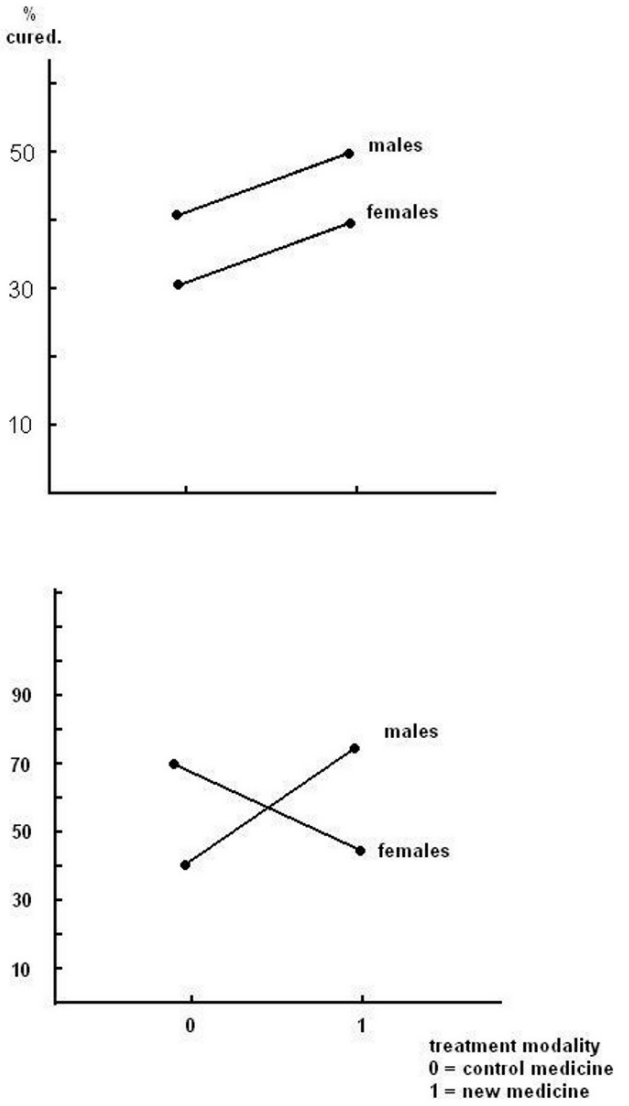


Figure 1. The effect of gender on a treatment comparison of two parallel groups treated with a new and a control medicine. Upper graph: the males respond better to both of the treatment than do the females, but no gender x treatment interaction is in the data. Lower graph: the data from the example given in the text: there is evidence for gender x treatment interaction because the males respond better to the new medicine, while the females respond better to the control treatment.

3. HOW TO TEST INTERACTION STATISTICALLY, A REAL DATA EXAMPLE WITH A CONCOMITANT MEDICATION AS INTERACTING FACTOR, INCORRECT METHOD

In the above example the presence of interaction is suggested. We can statistically test whether the difference between new medicine and control is significantly different using e.g. a chi-square or an odds ratio test. The tests produce p-values of >0.10 for the females and <0.001 for the males. It is tempting to state that the difference in p-values establishes a difference between the females and the males. P-values are composite estimators of not only effect size, but also spread in the data. Differences in p-values can arise because of differences in effect sizes or standard errors.

Table 1. Difference between females and males in odds ratios of treatment success.

Odds ratio treatment success females	standard error	p-value
0.658	1.336	>0.10
Odds ratio treatment success males		
3.667	1.357	<0.001
Difference in odds ratios[x]		
5.573	2.563	0.05<p<0.10

[x]The difference in odds ratios is calculated by subtracting their logarithmic transformations and turning this subtraction sum into its antilogarithmic term.

The correct approach is to compare directly the effect sizes relative to the standard errors, e.g. by using the odds ratios of treatment success in either subgroup (table 1). This procedure produces a p-value between 0.05 and 0.10. Obviously, there is a tendency to interaction. However, definitive evidence is lacking.

4. THREE ANALYSIS METHODS

	VERAPAMIL	METOPROLOL
MALES	52	28
	48	35
	43	34
	50	32
	43	34

	44	27	
	46	31	
	46	27	
	43	29	
	49	25	
	464	302	766
FEMALES	38	43	
	42	34	
	42	33	
	35	42	
	33	41	
	38	37	
	39	37	
	34	40	
	33	36	
	34	35	
	368	378	746
	832	680	

As an example we give a study of treatments for paroxysmal atrial fibrillation, the number of episodes per patient is the outcome variable.
Overall metoprolol seems to perform better. However, this is only true only for one subgroup (males).

First method, t-test

	Males	Females
Mean $_{verapamil}$ (SD)	46.4 (3.23866)	36.8 (3.489667)
Mean $_{metoprolol}$ (SD)	30.2 (3.48966) -	37.8 (3.489667) -
Difference means (SE)	16.2 (1.50554)	-1.0 (1.5606)

Difference of males and females 17.2 (2.166)

t-value = 17.2 / 2.166 = 8….

p <0.0001

There is a significant difference between the males and females, and, thus, a significant interaction between gender and treat-efficacy.

Second method, analysis of variance (ANOVA)

ANOVA Assesses whether the variance due to interaction is large compared to the variance due to chance (residual variance), (SS = sum of squares).

	verapamil	metoprolol
Males	52	28

	48	35	
	43		
	50	.	
	464 +	302 +	
Females	38	.	
	42	.	
	.	.	
	.		
	.	35	
	368 +	378+	746+
	832	680	1512

$SS_{total} = \frac{52^2 + 48^2 + \ldots 35^2}{} - \frac{(52+ 48+ + \ldots 35)^2}{40} = 1750.4$

$SS_{treatment\ by\ gender} = \frac{464^2 + \ldots 378^2}{10} - \frac{(52+ 48+ + \ldots 35)^2}{40} = 1327.2$

$SS_{residual} = SS_{total} - SS_{treatment\ by\ gender} = 423.2$

$SS_{rows} = \frac{766^2 + 746^2}{20} - \frac{(52+ 48+ + \ldots 35)^2}{40} = 10.0\ (= SS_{gender})$

$SS_{columns} = \frac{832^2 + 680^2}{20} - \frac{(52+ 48+ + \ldots 35)^2}{40} = 577.6\ (= SS_{treatment})$

$SS_{interaction} = SS_{treatment\ by\ gender} - SS_{rows} - SS_{columns} = 1327.2 - 10.0 - 577.6 = 739.6$

ANOVA-table (dfs = degrees of freedom, MS = mean square, F = F-statistic)

	SS	dfs	MS	F	P
Rows	10.0	1	10	0.851	ns
Columns (treatment)	577.6	1	577.6	49.1	<0.0001
Interaction	739.6	1	739.6	62.9	<0.0001
Residual	423.2	36	11.76		
Total					

In the above analysis the $SS_{interaction}$ is compared to the $SS_{residual}$. Often it is a better approach to use a "random-effects-model". The $SS_{treatment}$ is then compared to the $SS_{interaction}$. A p-value >0.05 indicates no interaction. Random effects models will be discussed in the chapter 38.

Third method, regression analysis

The y-variable is dependent, the x-variables are independent.

y = number of episodes of paroxysmal atrial fibrillation
x_1 = treat-modality (0 of 1)
x_2 = gender (0 of 1)
Add an additional interaction variable $x_3 = x_1 \times x_2$
Perform a multiple linear regression analysis including x_3 .

Regression-coefficients-table (b = regression coefficient)

	b	SE	t	sig
Constant	46.40	1.084	42.79	0.00
x_1	-16.20	1.533	-10.565	0.00
x_2	-9.60	1.533	-6.261	0.00
x_3 (interactie)	17.20	2.168	7.932	0.00

The t-value for x_3 = 7.932. The F-value for interaction in the above ANOVA-model = 62.916. It is interesting to observe that this F-value equals t^2 of the regression model. The two approaches are obviously very similar. We should note that for random-effects-modeling the SPSS software for linear-regression analyses has limited possibilities.

5. USING A REGRESSION MODEL FOR TESTING INTERACTION, ANOTHER REAL DATA EXAMPLE

How do we statistically test for the presence of interaction. Univariate analyses comparing subgroups can be used for that purpose. However, the linear regression model provides better sensitivity because it suffers less from missing data, and enables to analyze all of the data simultaneously. An example is provided by the Regress trial, a randomized parallel group trial of 884 patients treated with pravastatin or placebo for two years. The data of this study have already been briefly addressed in the chapters 12 and 14.[3] One of the primary efficacy variables was the decrease of the diameter of the coronary arteries after two years of treatment. The average decrease was 0.057 mm (standard error (SE) 0.013) in the pravastatin group, and it was 0.117 mm (SE 0.015) in the placebo group (t-test: significance of difference at $p<0.001$) (Figure 2, upper graph); thus the efficacy estimate b_1 was 0.060 (standard error SE = 0.016). Calcium antagonists had been given to 60% of the placebo patients, and to 59% of the pravastatin patients (chi-square: $p = 0.84$): thus, calcium antagonist treatment was not a confounder variable. Also, calcium antagonist medication was not associated with a diameter decrease ($p = 0.62$). In the patients who did not receive concomitant calcium antagonist medication, the diameter decreases were 0.097 (SE 0.014) and 0.088 (SE 0.014) in patients receiving placebo and pravastatin, respectively ($p = 0.71$). In patients who did receive calcium antagonist medication, the diameter decreases were 0.130 (SE 0.014) and 0.035 (SE 0.014), respectively ($p<0.001$). Thus, pravastatin - efficacy was, on average, 0.097-0.088 = 0.009 mm in the patients without calcium antagonist medication, and 0.130-0.035 = 0.095 in the patients with calcium antagonist medication (Figure 2, lower graph). The two lines cross, suggesting the presence of interaction between pravastatin and calcium antagonists.

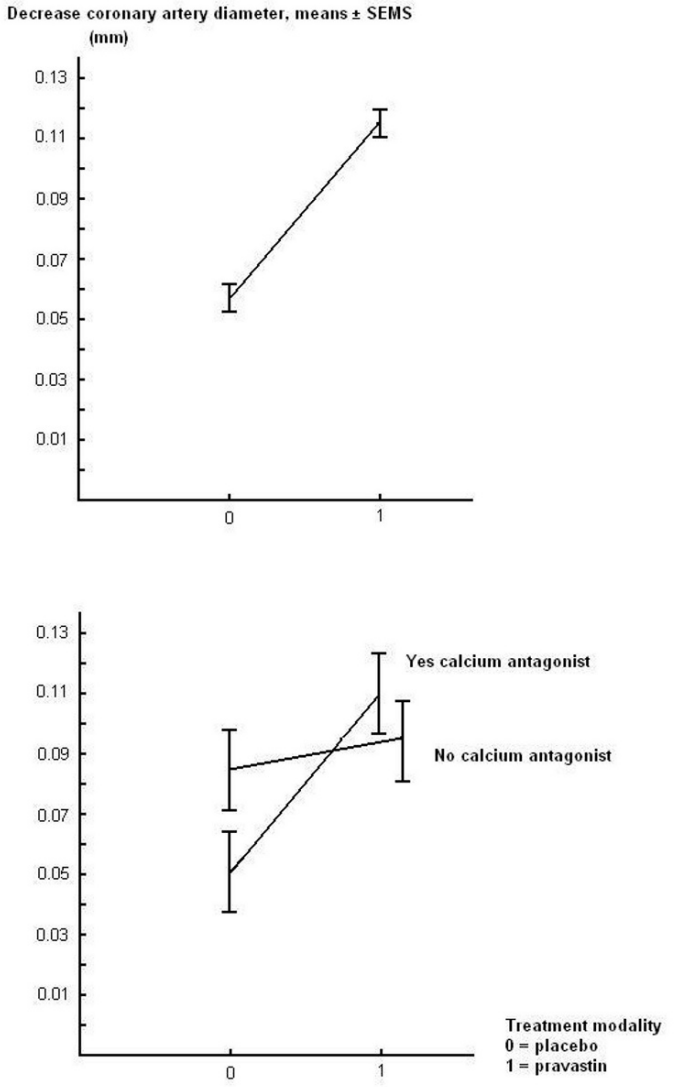


Figure 2. The effect of concomitant calcium antagonists on treatment efficacy of pravastatin estimated by the decrease of coronary artery diameter (ca-diameter) after two years' treatment (REGRESS data[4]). Upper graph: pravastatin significantly decreased ca-diameter compared to placebo. Lower graph: there is evidence for interaction between calcium antagonists and pravastatin, because in the patients receiving no calcium antagonist the benefit of pravastatin was insignificant, while it was highly significant in the patients receiving a concomitant calcium antagonist.

Before statistically testing this suggested interaction, we have to assess whether it makes clinically sense. Atherosclerosis is characterized not only by depots of cholesterol but also of calcium in the fatty streaks that consist of foam cells. It does make sense to argue that calcium antagonists, although they do not reduce plasma calcium, reduce calcium levels in the foam cells, and, thus, beneficially influence the process of atherosclerosis, and that interaction with cholesterol lowering treatment is a possibility.
We used the following linear regression model for this test:

$y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + e_i$
where
y_i = dependent variable = decrease in coronary artery diameter in the ith patient
a = intercept
b_1 , b_2, and b_3 = partial regression coefficients for the variables (1) treatment modality, (2) calcium antagonist treatment, (3) interaction between (1) and (2).
e_i = systematic error in the ith patient

Let x_{1i}=1 denote that patient i received pravastatin (x_{1i}=0, if not), let x_{2i}=1 denote that patient i received calcium antagonist medication (x_{2i} = 0, if not), and let x_{3i} = x_{1i} times x_{2i} . The estimates were: b_3 = 0.085 (SE 0.033), b_2 = -0.033 (SE 0.023), and b_1 = 0.009 (SE 0.026). Notice that b_1 changed dramatically by including the interaction term x_3 in the linear model; this is a general feature of regression models with interaction terms: the corresponding main-effects (b_1 and b_2) cannot be interpreted independently of the interaction term. Another consequence is that the efficacy estimate no longer exists, but several estimates do exist: in our case there are different efficacy-estimates for patients with (b_1+b_3 = 0.009+0.085 = 0.094) and without calcium antagonist medication (b_1 = 0.009). This difference was statistically significant (interaction test: p=0.011).

6. ANALYSIS OF VARIANCE FOR TESTING INTERACTION, OTHER REAL DATA EXAMPLES

Parallel-group study with treatment x health center interaction
Current clinical trials of new treatments often include patients from multiple health centers, national and international. Differences between centers may affect results. We might say these data are at risk of interaction between centers and treatment efficacy. Hays[3] described an example: 36 patients were assessed for performance after treatment with either placebo, vitamin supply low dose, and high dose. Patients were randomly selected in 6 health centers, 6 patients per center, and every patients was given one treatment at random, and so in each center two patients were given one of the three treatments. The Table 1 gives an overview of the results.

Table 1. Results of three treatments for assessment of performance in 36 patients in 6 health centers, results are given as scores (data modified from Hays[3] with permission from the editor)

Treatment	placebo	vitamin supply low dose	high dose	total
Health center				
1	7.8	11.7	11.1	
	8.7	10.0	12.0	
	16.5	21.7	23.1	61.3
2	8.0	9.8	11.3	
	9.2	11.9	10.6	
	17.2	21.7	21.9	60.8
3	4.0	11.7	9.8	
	6.9	12.6	10.1	
	10.9	24.3	19.9	55.1
4	10.3	7.9	11.4	
	9.4	8.1	10.5	
	19.7	16.0	21.9	57.6
5	9.3	8.3	13.0	
	10.6	7.9	11.7	
	19.9	16.2	24.7	60.8
6	9.5	8.6	12.2	
	9.8	10.5	12.3	
	19.3	19.1	24.5	62.9
total	103.5	119.0	136.0	358.5

The model is $y = \mu + a + b + ab + e$

where y = dependent variable, estimate for performance of patients
- μ = mean result
- a = fixed effect of the three treatments
- b = random variable associated with health center
- ab = random interaction effect between treatment and health center
- e = systematic error

The computations are (SS= sum of squares)

$$SS_{total} = (7.8)^2 + \ldots + (10.5)^2 - \frac{(358.5)^2}{36} = 123.56$$

$$SS_{ab} = \frac{(16.5)^2 + \ldots.. + (19.1)^2}{2} - \frac{(358.5)^2}{36} = 109.03$$

$$SS_{error} = SS_{total} - SS_{ab}$$
$$= 123.57 - 109.03 = 14.54$$

$$SS_{columns} = \frac{(103.5)^2 + (136.0)^2 + (119.0)^2}{12} - \frac{(358.5)^2}{36} = 44.04$$

$$SS_{rows} = \frac{(61.3) + \ldots.. + (62.9)^2}{6} - \frac{(358.5)^2}{36} = 6.80$$

$$SS_{interaction} = SS_{ab} - SS_{rows} - SS_{columns}$$
$$= 109.03 - 6.80 - 44.04 = 58.19$$

Table 2 gives the ANOVA (analysis of variance) table. The F test for interaction

Table 2. ANOVA table of analysis for data of Table 1

Source	SS	dfs	MS	F
Columns	44.04	3-1=2	22.02	22.02/5.82=3.78
Rows (centers)	6.80	6-1=5	1.36	1.68
Interaction (treatment x center)	58.19	10	5.82	5.82/0.81=7.19
Error	14.54	18x(2-1)=18	0.81	
Total	123.57	35		

SS = sum of squares
dfs = degrees of freedom
MS = mean square
F = test statistic for F test

Produces an F-value of 7.19 corresponding with a p-value <0.01 which means that the hypothesis of no interaction is rejected. Although there is insufficient evidence to permit to conclude that there are treatment effects or health center effects, there is pretty strong evidence for the presence of interaction effects. There is something about the combination of a particular health center with a particular treatment that accounts for a significant part of the variability in the data. Thus, between the health centers, treatment differences apparently exist. Perhaps the capacity of a treatment to produce a certain result in a given patient depends on his/her health center background.

Crossover study with treatment x subjects interaction

In a crossover study different treatments are assessed in one and the same subject. Suppose, we have prior arguments to believe that subjects who better respond to one treatment, will also do so to another treatment. E.g., in trials involving a similar class of drugs subjects who respond better to one drug, tend to respond better to all of the other drugs from the same class, and those who respond less, will respond less to the entire class. For example, patients with angina pectoris, hypertension, arrhythmias, chronic obstructive pulmonary disease, responsive to one class of drugs may equally well respond to a different compound from the same class. In this situation our interest may focus on the question is there a difference in response between different patients, instead or in addition to the question is there an overall difference between treatments. If the emphasis is on the differences between the subjects, the design is often called a treatments - by - subjects design. An example is in Table 3.

Table 3. Diastolic blood pressures (mm Hg) after 4 week treatment with four different treatments in a crossover study of 12 patients

Patient	Treatment 1	treatment 2	treatment 3	treatment 4	sd^2
1	98	96	98	90	…
2	94	92	92	86	…
3	92	94	94	88	
4	94	90	90	90	
5	96	98	98	96	
6	94	88	90	92	
7	82	88	82	80	
8	92	90	86	90	
9	86	84	88	80	
10	94	90	92	90	
11	92	90	90	94	
12	90	80	80	80	
	1104	1080	1080	1056	Add-up sum = 4320

Table 4. ANOVA table for the data of Table 3

Source	SS	dfs	MS	F	p-value
Subjects	906	12-1=11			
Treatments	96	4-1=3	8	8/1.74=4.60	<0.05
Subjects x treatments	230	3x11=33			
Total	1232	47			

SS = sum of squares
dfs = degrees of freedom
MS = mean square
F = test statistic for F test

Twelve patients are given in random order 4 different antihypertensive drugs from the same class. Diastolic blood pressures were used as variable. The statistical model with computations are (sd = standard deviation):

$SS_{subjects} = sd_1^2 + sd_2^2 + \ldots\ldots sd_{12}^2 = 906$
$SS_{treatments}$ = (treatment mean 1 – grand mean)^2 + (treatment mean 2 – grand mean)^2 +…= 96.0

$$SS_{total} = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = (49^2 + \ldots\ldots 40^2) - (2160)^2 / 48 = 1232$$

$SS_{subjects\ x\ treatments} = SS_{total} - SS_{subjects} - SS_{treatments} = 230$

The layout for this repeated measures situation is given in Table 4. The MS (mean square) for treatments divided by the MS for subjects - by - treatments interaction gives an F - ratio of 4.60. If we are using an alpha level of 0.05 for this test, this results will be significant. The four treatments appear to be having different effects to different subsets in this sample. Note that an overall F test on these data requires the SS residual term which is equal to SS subjects – SS treatments. The F - ratio used for an overall F test equals MS treatments / MS residual, and would produce an entirely different result (see also chapter 2).
From the above analysis it can be concluded that an interaction effect exists between treatments and patients. Some patients, obviously, respond better or worse to the treatments than others. This is probably due to personal factors like genetic polymorphisms, societal and/or developmental factors. This repeated measures model is particularly convenient in drug development that has to account such factors when assessing the elimination rate and other pharmacokinetic properties of new drugs. Statistical models like these are often called mixed effects models, because they are considered to include a fixed effect (the treatment effect), and a

random effect (the effect of being in a subset). Mixed effects models will be further discussed in chapter 29.

7. DISCUSSION

Interaction effects in a clinical trial should be distinguished from confounding effects. In a trial with interaction effects the treatment groups are generally nicely symmetric. However, there are subsets in each treatment group that have an unusually high or low response. With confounding, things are different. For whatever reason the randomization failed, and the treatment groups are different for a clinically relevant factor. E.g., in a placebo-controlled trial the two parallel-groups were asymmetric for age. The control group was significantly older than the treatment group, and this could easily explain the treatment difference. More examples of confounding are given in the chapters 12 and 16.

Also interaction effects should be distinguished from carryover effects as commonly observed in crossover studies, and sometimes wrongly called treatment by period interaction. If in a crossover study the effect of the first period of treatment carries on into the second period of treatment, then it may influence the response to the latter period. More examples of this phenomenon will be given in the chapters 19 and 20.

Clinical trials usually do not include interaction assessments in the protocol. Results of such assessments are, therefore, post-hoc, and of an exploratory and unconfirmed nature. Why should they be performed even so? In cardiovascular research drug-drug interactions, and effects of comorbidities on drug efficacies are numerous. It is valuable to account at least post-hoc for such mechanisms. Second, current clinical trials involve heterogeneous health centers, investigators, and patient groups. Accounting these heterogeneities can be helpful to predict individual responses in future patients, and to develop prediction rules based on the trial data of individuals. Prediction rules like the Framingham risk score may be developed using trial data to further identify subjects at risk of having a good or bad drug response.

Other reasons for interaction assessments include the following. It may be useful and reassuring to know that in a positive study the benefit in subgroups parallels the benefit in the study overall. E.g., in a subgroup analysis of the Dietary Approach to Stop-Hypertension (DASH) randomized clinical trial the results were equivalent in different age, gender, and ethnic groups.[5] Also, it may be useful to know if there are in an unexpectedly negative study certain subgroups that might be benefited or harmed by the treatment. E.g., estrogen / progestin replacement caused cardiovascular benefit in women with high lipoprotein, harm in those with low lipoprotein.[6]

We should add that the assessment of interaction, otherwise called heterogeneity, is not always wise. In controlled clinical trials a myriad of subgroups can be identified that would qualify for an exploratory examination, and this approach will almost certainly produce one or more spuriously significant interactions. Interaction terms to be assessed should, therefore, make clinically sense. Demonstrating a statistically

significant interaction between the treatment effect and the first letters the patients' Christian names makes no sense, and pursuing such a finding is merely data dredging. We should caution that a "scientific" explanation can be found for every subgroup results in one afternoon Pubmed search. As stated by Dr. Barrett-Connor in a commentary on the 9 positive interactions demonstrated in a subgroup analysis of the otherwise negative Heart and Estrogen / progestin Replacement Study (HERS) trial, biological plausibility is quite easy to theorize, anyone with 2 hours and a little imagination can do it.[7]

Statistical methods for identifying and integrating interaction terms into the data are limited, and have a limited statistical power. Moreover, if we introduce a large number of interaction terms in a regression analysis, the statistical power to demonstrate statistical significance for the primary endpoint will be reduced. Nonetheless, the assessment of a small number of interaction terms in clinical research can be an important part of the evaluation of new drugs.

The issue of testing interaction is different with meta-analyses of clinical trials.[8] The aim of a meta-analysis is to obtain a pooled estimate of a treatment effect rather the study of subgroups. The studies to be included in a meta-analysis are often heterogeneous, and protocols, therefore, routinely apply a heterogeneity test prior to data pooling. In the presence of a statistically significant heterogeneity, an data pooling may be difficult to accept, and has as an additional problem that confidence intervals are underestimated, because the extra variability between the different trials is ignored.

If a statistically significant interaction is demonstrated post hoc, its existence should be confirmed in a novel prospective clinical trial. If a relevant interaction is expected prior to the trial, its assessment should be properly included in the trial protocol at the planning stage of the trial. Instead of a regression model a factorial trial design is suitable for such purposes.

Linear regression analyses may provide better precision to test interaction than comparison of subgroups.[3] It is also often more convenient, because it enables to analyze all of the data simultaneously. Different regression models may be adequate for different types of data, e.g., exponential models are more adequate than linear models for risk ratios and mortality data.

8. CONCLUSIONS

In pharmaceutical research and development, multiple factors co-determine the efficacy of the new treatment. In statistical terms we say they interact with the new treatment efficacy. Interaction effects, are not routinely assessed in clinical trials. The current paper reviews some important factors that may interact with the treatment efficacy, and comes to the following recommendations:

1. The assessment of a small number of interaction terms is an important part of the evaluation of new medicines. Important factors that may interact with the treatment efficacy are: (a) concomitant drugs and/or comorbidities, (b) health center factors in

multicenter trials, (c) subject factors like genetic polymorphisms relating to the speed of drug metabolism.
2. Interaction terms to be assessed should make clinically sense.
3. Linear regression analyses provide better sensitivity to test interaction than do subgroup analyses, because they suffer less from missing data and enable to analyze all of the data simultaneously. Exponential regression models are more adequate for risk ratios and mortality data.
4. If a relevant interaction is clinically expected, its assessment should be properly included in the trial protocol at the planning stage of the trial.
5. If a statistically significant interaction is demonstrated post hoc, its existence should be confirmed in a novel prospective clinical trial.
We hope that the examples and recommendations in this chapter be guidelines for the analysis of interaction effects in clinical drug trials, in order to better predict the efficacy / safety of new medicines in future clinical treatment of individual patients.

9. REFERENCES

1. Riegelman RK. Studying a study and testing a test. Lippincott Williams & Wilkins, Philadelphia, PA, 2005.
2. De Craen AJM, Westendorp RGJ. The use of age as a variable in clinical research. Ned Tijdschr Geneeskd 2005; 149: 2958-63.
3. Hays WL. Random effects and mixed models. Chapter 13. In: Statistics, 4th edition, Holt, Rhinehart and Winnston Inc, Chicago, 1998, pp 479-543.
4. Jukema AJ, Zwinderman AH, et al for the REGRESS Study Group. Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elevated serum cholesterol levels. The Regression Growth Evaluation Statin Study (REGRESS). Circulation 1995; 91: 2528-40.
5. Svetkey LP, Simons- Morton D, Vollmer WM, et al. Effects of dietary patterns on blood pressure: subgroup analysis of the Dietary Approaches to Stop -Hypertension (DASH) randomized clinical trial. Arch Int Med 1999; 159: 258 -93.
6. Shlipak MG, Simon JA, Vittinghoff E, et al. Estrogen and progestin, lipoprotein (a), and the risk of recurrent coronary heart disease events after menopause. JAMA 2000; 283: 1845-52.
7. Barrett-Connor E. Looking for the pony in the Heart and Estrogen / progestin Replacement Study (HERS) data. Circulation 2002; 105: 902-3.
8. Chalmers I, Altman DG. Systematic reviews. Edited by Br Med J Books, Bristol Uk, 1996.

CHAPTER 21

META-ANALYSIS, BASIC APPROACH

1. INTRODUCTION

Problems with meta-analyses are frequent: regressions are often nonlinear; effects are often multivariate rather than univariate; continuous data frequently have to be transformed into binary data for the purpose of comparability; bad studies may be included; coverage may be limited; data may not be homogeneous; failure to relate data to hypotheses may obscure discrepancies. In spite of these well-recognized flaws, the method of meta-analysis is an invaluable scientific activity: Meta-analyses establish whether scientific findings are consistent and can be generalized across populations and treatment variations, or whether findings vary significantly between particular subsets. Explicit methods used limit bias and improve reliability and accuracy of conclusions, and increase the power and precision of estimates of treatment effects and risk exposures. In the past decade, despite reservations on the part of regulatory bodies, the method of meta-analysis has increasingly been employed in drug development programs for the purpose of exploration of changes in treatment effect over time, integrated summaries of safety and efficacy of new treatments, integrating existing information, providing data for rational decision making, and even prospective planning in drug development.

Meta-analyses are increasingly considered an integral part of phase III drug research programs for two reasons. First, meta-analysis of existing data instead of an unsystematic literature search before starting a phase III drug trial has been documentedly helpful in defining the hypothesis to be tested. Second, although meta-analyses are traditionally considered post-hoc analyses that do not test the primary hypotheses of the data, they do test hypotheses that are extremely close to the primary ones. It may be argued, therefore, that with the established uniform guidelines as proposed by Oxman and Guyatt and implemented by the Cochrane Collaborators, probability statements are almost as valid as they are in completely randomized controlled trials.

Meta-analyses should be conducted under the collective responsibility of experienced clinicians and biostatisticians familiar with relevant mathematical approaches. They may still be improved, by a combination of experience and theory, to the point at which findings can be taken as sufficiently reliable where there is no other analysis or confirmation is available.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 263–276.

Meta-analyses depend upon quantity and quality of original research studies as reported. Helpful initiatives to both ends include the Unpublished Paper Amnesty Movement endorsed by the editors of nearly 100 international journals in September 1997 which will help to reduce the quantity of unpublished papers, and the Consolidated Standards of Reporting Trials (CONSORT) Statement (1997) developed by high impact journals which is concerned with quality and standardization of submitted papers.

Meta-analysis can help reduce uncertainty, prevent unnecessary repetition of costly research, and shorten the time between research discoveries and clinical implementation of effective diagnostic and therapeutic treatments, but it can only do so when its results are made available. The continuously updated Cochrane Database of Systematic Reviews on the Internet is an excellent example for that purpose. Medical journals including specialist journals have a responsibility of their own. So much so that they may be able to lead the way for biased experts, who are so convinced of their own biased experience and so little familiar with meta-analysis.

2. EXAMPLES

We have come a long way since psychologists in the early 70s drew attention to the systematic steps needed to minimize biases and random errors in reviews of research. E.g., we currently have wonderful meta-analyses of pharmacological treatments for cardiovascular diseases which helped us very much to make proper listings of effective treatments (as well as less effective ones). So, now we are able to answer 1st what is best for our patients, 2nd how we should distribute our resources. For example, for acute myocardial infarction, thrombolytic therapy as well as aspirin are highly effective, while lidocaine and calcium channel blockers are not so. For secondary prevention myocardial infarction cholesterol-reducing therapy were highly effective while other therapies were less so or was even counterproductive, e.g., class I antiarrhythmic agents as demonstrated in Figure 1.
On the x-axis we have odds ratios. Many physicians have difficulties to understand the meaning of odds ratios. Odds = likelihood = chance = probability = risk that an event will occur divided by the chance that it won't. It can be best explained by considering a four cell contingency table.

Contingency table	numbers of subjects who died	numbers of subjects who did not die
Test treatment (group$_1$)	a	b
Control treatment (group$_2$)	c	d

The proportion of subjects who died in group1 (or the risk (R) or probability of having an effect)

= p = a / (a+b) , in group 2 p= c / (c+d),
the ratio of a / (a+b) and c / (c+d) is called risk ratio (RR) .

Another approach is the odds approach, where a/b and c/d are odds, and their ratio is the odds ratio (OR). In meta-analyses of clinical trials we use ORs as surrogate RRs, because, here, a / (a+b) is simply nonsense.

For example:

	treatment group(n)		control group(n)		whole population(n)
Sleepiness(n)	32	a	4	b	4000
No sleepiness(n)	24	c	52	d	52000

n= numbers of patients.

We assume that the control group is just a sample from the population, but its ratio, b/d, is that of the population. So, suppose 4 = 4000, and 52 = 52000, then the term $\frac{a/(a+b)}{c/(c+d)}$ suddenly becomes close to the term $\frac{a/b}{c/d} = RR$ of the population.

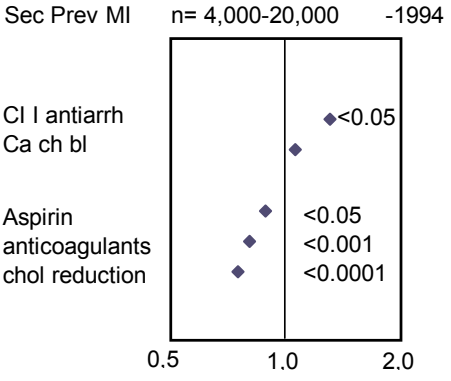


Figure 1. Pooled results (odds ratios = odds of infarction in treated subjects / odds of infarction in controls) of secondary prevention trials heart infarction.

Currently, even epidemiologists are borrowing from clinical pharmacologists and clinical investigators, and they are quite successful in showing the likeliness of various epidemiological issues such as the epidemiology of various cardiovascular

conditions. It should be emphasized that the logic behind meta-analysis is simple and straightforward. All it requires, is to stick to the scientific methods, that is (1) a clearly defined prior hypothesis, (2) thorough search of trials, (3) strict inclusion criteria for trials, and (4) uniform guidelines for data analysis.

3. CLEARLY DEFINED HYPOTHESES

In chapter 1 we discussed that drug trials principally address efficacy and safety of new drugs. It is specified in advance –in the statistical analysis plan- what are the main outcome variables, and how they should be tested.
A meta-analysis is very much similar to a single trial, and similarly to a single trial it tests a very small number of primary hypotheses, mostly the hypotheses that the new compound is more efficaceous and safe than the reference compound. This implies that data dredging is as unacceptable for meta-analyses as it is for separate clinical trials.

4. THOROUGH SEARCH OF TRIALS

The activity of thoroughly searching-published-research requires a systematic procedure. E.g., searching medline requires a whole lot of tricks, and has to be learned. Unless you already know, you may pick up a checklist for this purpose, similarly to the checklist used by aircraft staff before take off, a nice simile used by Dr Oxman from McMasters University, one of the enlightened specialists of meta-analyses. A faulty review of trials is as perilous as a faulty aircraft and both of them are equally deadly, the former particularly so if we are going to use it for making decisions about health care. Search terms will soon put you on the right track when searching Medline. SH, e.g., means “subject-heading” which is controlled vocabulary; TW means “free-text-word” (searching with a lot of TWs increases sensitivity but reduces specificity of the search. There are sensitive ways to look for RCTs. ADJ is another TW and is more precise than AND. NOT means that first and third step are combined and second step is excluded. Use of checklists consistent of search terms of controlled vocabulary and frequent use of free text words makes things so much easier and overcomes the risk of being unsuccessful.

5. STRICT INCLUSION CRITERIA

The third scientific rule is strict inclusion criteria. Inclusion criteria are concerned with validity of the trials to be included, which means their likeliness of being unbiased. Strict inclusion criteria means that we subsequently only include the valid studies. A valid study is an unbiased study , a study that is unlikely to include systematic errors. The most dangerous errors in reviews are systematic errors otherwise called biases. Checking validity is thus the most important thing both for doers and for users of systematic reviews. Some factors have empirically been shown to beneficially influence validity. These factors include: blinding the study;

random assignment of patients; explicit description of methods; accurate statistics; accurate ethics including written informed consent.

6. UNIFORM DATA ANALYSIS

Statistical analysis is a tool which, when used appropriately, can help us to derive meaningful conclusions from the data. And it can help us to avoid analytic errors. Statistics should be simple and should test primary hypotheses in the first place. Before any analysis or plotting of data can be performed we have to decide what kind of data we have.

1. Individual data

Primary data of previously published studies are generally not available for use. Usually, we have to accept the summary statistics from studies instead. This is of course less informative and less precise than a synthesis of primary data but can still provide useful information.

2. Continuous data, means and standard errors of the mean (SEMs)

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

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

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

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

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

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

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

3. Proportions: relative risks (RRs), odds ratios (ORs), Differences between relative risks (RDs)

Probably, 99% of meta-analyses make use of proportions rather than continuous data, even if original studies provided predominantly the latter particularly for efficacy data (mean fall in blood pressure etc.). This is so both for efficacy and safety meta-analyses. Sometimes data have to be remodeled from quantitative into binary ones for that purpose.

Calculation of point estimates and their variances

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

Point estimators RR, OR, or RD:

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

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

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

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

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

$$\text{Chi-square} = \frac{\left(\frac{\text{lnRR}_1}{s_1^2} + \frac{\text{lnRR}_2}{s_2^2} + \frac{\text{lnRR}_3}{s_3^2} \ldots\right)^2}{\frac{1}{s_1^2} + \frac{1}{s_2^2} + \frac{1}{s_3^2} + \ldots} \quad \text{degrees of freedom 1 (one).}$$

s^2 = variance of point estimate :
$s_{\text{lnRR}}{}^2 = 1/a - 1/(a+b) + 1/c - 1/(c+d)$
$s_{\text{lnOR}}{}^2 = 1/a + 1/b + 1/c + 1/d$
$s_{\text{RD}}{}^2 = ab/(a+b)^3 + cd/(c+d)^3$

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

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

As alternative approach Mantel-Haenszl-summary chi-square can be used:

Mantel-Haenszl summary chi-square test:

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

a_i, b_i , c_i, and d_i are the a-value, b-value, c-value, and d-value of the ith sample

This approach has been explained in chapter 3. Results of the two approaches yield similar results. However, with Mantel-Haenszl the calculation of pooled variances is rather complex, and a computer program is required.
A good starting point with any statistical analysis is plotting the data (Figure 2).

4. Publication bias

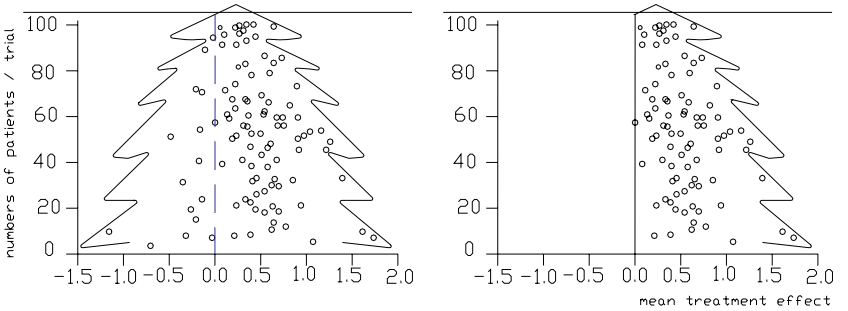


Figure 2. This Christmas tree otherwise called funnel plot of 100 published trials shows on the x-axis the mean result of each trial; on the y-axis it shows the numbers of pts involved in each trial. As you can see on the left, there is a Christmas-tree or upside-down-funnel-pattern of distribution of the results. The smaller the trial, the larger the distribution of results. Right graph gives a simulated pattern, suggestive for ***publication bias****: the negative trials are not published and thus missing. This cut Christmas-tree can help us suspect that there is a considerable publication bias in the meta- analysis.*

This socalled funnel plot of 100 published trials shows on the x-axis the mean result of each trial; on the y-axis it shows the numbers of pts involved in each trial. As you can see on the left, there is a Christmas-tree or upside-down-funnel-pattern of distribution of the results. The smaller the trial, the larger the distribution of results. Right graph gives a simulated pattern, suggestive for publication bias: the negative trials are not published and thus missing. This cut Christmas-tree can help us suspect that there is a considerable publication bias in the meta-analysis. **Publication bias** can also be statistically tested by rank correlation between variances and odds ratios. If small studies with negative results are less likely to be published, rank correlation would be high, if not it would be low. This can be assessed by the **Kendall tau test**:

Normally, the correlation coefficient r measures actual results. The Kendall tau-test basically does the same, but uses ranked data instead of actual data.

Trial	A	B	C	D	E	F	G	H	I	
Ranknumber of size of trial	1	2	3	4	5	6	7	8	9	10
Ranknumber of size of mean result	5	3	1	4	2	7	9	6	10	8

Lower row add up rank numbers higher than 5, respectively 3, respectively 1, respectively 4: we find 5+6+7+5+5+3+1+2+0+0=34.
Then lower row add up rank numbers lower than 5, 3, 1, etc: we find 4+2+0+1+0+1+2+0+1+0=11.

The standard error of this result is $\sqrt{\frac{n(n-1)(2n+5)}{18}}$, and we assume a normal distribution. We can now test this correlation and find

$$\frac{(34-11)}{\sqrt{\frac{n(n-1)(2n+5)}{18}}} = 1.968$$

which is approximately 1.96 = 2 = the number of SEMs distant from which is ≤5% of the data. And so, the null-hypothesis of no publication bias has to be rejected. **Publication bias** can also be tested by calculating the shift of odds ratios caused by the addition of unpublished trials e.g. from abstract-reports or proceedings.

5. Heterogeneity

Figure 3 gives an example of a meta-analysis with means and 95% confidence intervals (CIs), telling us something about heterogeneity.

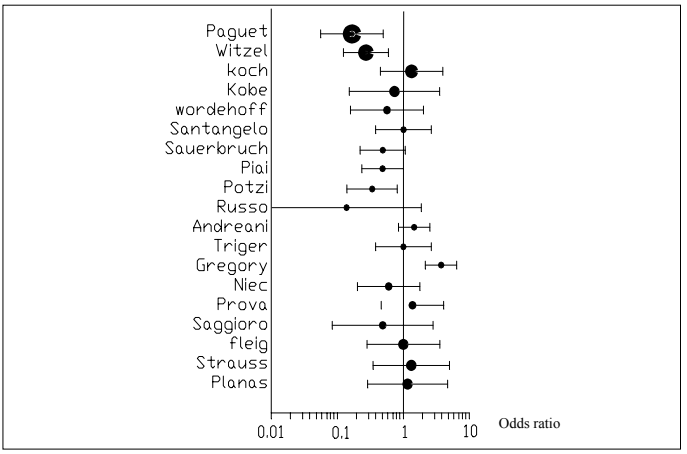


Figure 3. Heterogeneous trials, 19 trials of endoscopic intervention vs no intervention for upper intestinal bleeding. On the y-axis the individual studies, on the x-axis the results, the sizes of the bullets correspond to the sizes of the studies (with permission from the editor).[1]

On the x-axis is the result, on the y-axis are the trials. This example has been previously used by Dr Thompson from London School of Hygiene and Tropical Medicine. We see the results of 19 trials of endoscopic sclerotherapy for esophageal varices bleeding: odds ratios less than one represent a beneficial effect. These trials were considerably different in patient-selection, baseline-severity-of-condition, sclerotechniques, management-of-bleeding-otherwise, and duration-of-follow-up.

And so, this is a meta-analysis which is clinically very heterogeneous. Is it also statistically heterogeneous? For that purpose we test whether there is a greater variation between the results of the trials than is compatible with the play of chance, simply using a chi-square test. In so doing, we find χ^2= 43 for 19-1=18 degrees of freedom (dfs). The p value is < 001 giving substantial evidence for statistical heterogeneity. For the interpretation of such tests it is useful to know that a χ^2 statistic has on average a value equal to the degrees of freedom, so a result of χ^2= 18 with 18 dfs would give no evidence for heterogeneity, values much larger such as here observed do so for the opposite.

With very few studies in the meta-analysis, or with small studies, the fixed model approach has little power, and is susceptible to type II errors of not finding heterogeneity which may actually be in the data. A little bit better power is then provided by the random effect model of Dersimonian and Laird, which assumes an additional variable. The variable $s_{\text{between trials}}$ is added to the model, meaning the size of variance between the trials. The fixed model for testing the presence of heterogeneity of ordinal data is demonstrated underneath. For continuous data multiple group analysis of variance (ANOVA) may be used).

Fixed effect model (Cochran-Q test)
test for homogeneity (k-1 degrees of freedom)

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

Random effect model (DerSimonian and Laird)
Test for heterogeneity is identical, except for variances s^2 which are replaced with $(s^2 + s_{\text{between trials}}^2)$.

Example of Random effect model analysis

trial	test treatment		reference treatment	
	deaths	survivors	deaths	survivors
1	1	24	5	20
2	5	95	15	85
3	25	475	50	450

In the above example, the test for heterogeneity fixed effect model provides χ^2 = 1.15 with dfs 3-1 = 2, while the test with the random effect model provides a χ^2 = 1.29 with dfs equally 2, both lower than 2. The between-trial variance $s_{\text{between trials}}^2$ is thus accepted to be zero and the weights of the two models are equal. Heterogeneity can be neglected. With the simple example given, the two approaches to test homogeneity raise similar results (the null

hypothesis is tested that studies are equal). And so, between-trial variance $s_{between\ trials}^2$ is accepted to be zero and the results of the two models are equal.
Heterogeneity can be neglected in this example.

Heterogeneity and sub-group analysis

When there is heterogeneity, to analists of systematic reviews, that's when things first get really exciting. A careful investigation of the potential cause of heterogeneity has to be accomplished. The main focus then should be on trying to understand any sources of heterogeneity in the data. In practice, this may be less hard to assess since the doers have frequently noticed clinical differences already, and it thus becomes relatively easy to test the data accordingly. Figure 4 below shows how age e.g. is a determinant of illness, but in the right graph the risk difference is heterogeneous because it increases with age.

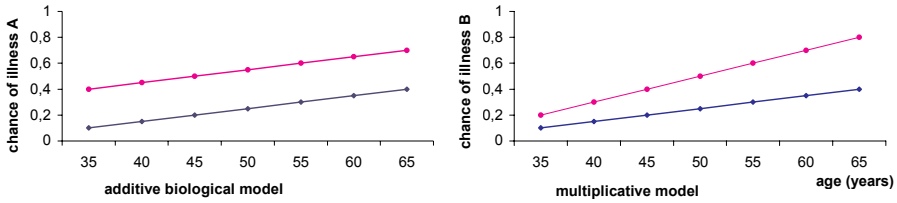


Figure 4. Age is a determinant of illness, but in the right graph the risk difference is heterogeneous because it increases with age.

Except age, outliers may give an important clue about the cause of heterogeneity.

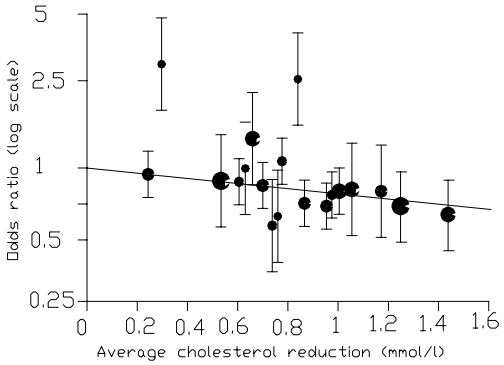


Figure 5. The relation between cholesterol and coronary heart disease. The two outliers on top were the main cause for heterogeneity in the data, the sizes of the bullets correspond to the sizes of the studies (with permission from the editor).[2]

Figure 5 shows the relation between cholesterol and coronary heart disease. The two outliers on top were the main cause for heterogeneity in the data: one study was

different because it achieved a very small reduction of cholesterol; the other was a very short-term study.
Still other causes of heterogeneity may be involved. 33 Studies of cholesterol and risk of carcinomas showed that heterogeneity was huge. When the trials were divided according to social class, the effect in the lowest class was 4 -5 times those of the middle and upper class, explaining everything about this heterogeneous result.
We should, of course, warn of the danger of overinterpretation of heterogeneity. Heterogeneity may occur by chance. This is particularly an important possibility to consider when no clinical explanation is found. Also, we should warn that a great deal of uniformity among the results of independently performed studies is not necessarily good; it can suggest consistency-in-bias rather than consistency-in-real-effects.

6. Robustness

Sensitivity or robustness of a meta-analysis is one last important aspect to be addressed in the analysis of the data. When talking of strict inclusion criteria, we discussed studies with lower levels of validity, as assessed by factors such as blinding, random assignments, accurate and explicit description of results and statistics. It may be worthwhile not to completely reject the studies with lower methodology. They can be used for assessing another characteristic of meta-analyses, namely its sensitivity.

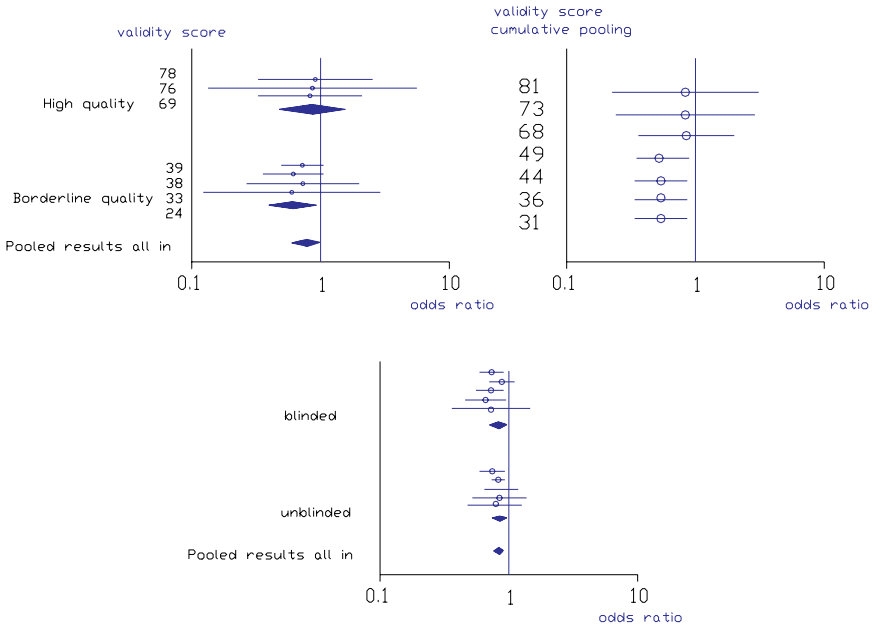


Figure 6. The left upper graph gives an example of how the pooled data of three high-quality-studies provide a smaller result, than do 4 studies-of-borderline-quality. The summary result is mainly determined by the borderline-quality-studies, as is also shown in the cumulative-right-upper-graph. When studies are ordered according to their being blinded or not as shown in the lower graph, differences may be large or may be not so. In studies using objective variables, e.g., blood pressures, heart rates, blinding is not so important than in studies using subjective variables (pain scores etc). In this particular example differences were negligible.

This left upper graph (Figure 6) gives an example of how the pooled data of three high-quality-studies provide a smaller result, than do 4 studies-of-borderline-quality. The summary result is mainly determined by the borderline-quality-studies, as is also shown in the cumulative-right-upper-graph. When studies are ordered according to their being blinded or not as shown in the lower graph, differences may be large or may be not so. In studies using objective variables, e.g., blood pressures, heart rates, blinding is not so important than in studies using subjective variables (pain scores etc). In this particular example differences were negligible. So, in conclusion, when examining the influence of various inclusion criteria on the overall odds ratios, we may come to conclude that the criteria themselves are an important factor in determining the summary result. We say in that case that the meta-analysis lacks robustness (otherwise called sensitivity or precision of point estimates). Interpretation then has to be cautious, pooling may have to be left out altogether. Just leaving out trials at this stage of the meta-analysis is inappropriate either, because it would introduce bias similar to publication-bias or bias-introduced-by-not-complying-with-the-intention-to-treat-principle.

7. DISCUSSION, WHERE ARE WE NOW?

Several recent publications were critical of the method of meta-analysis: e.g., Chalmers and Lau in JAMA 1996 and Lelorier in NEJM 1997 concluded that meta-analyses did not accurately predict the outcomes of subsequent large trials. Colditz and Berlin JAMA 1999 concluded that meta-analyses were not or at least not-yet good enough to identify adverse drug reactions. Why so? Probably, the answer is (1st) trials must get better, and (2nd) publication bias must disappear altogether. There are several important initiatives being taken at this very moment that may be helpful to this aim. In May 1998 editors of 70 journals have endorsed the Consolidated- Standards-of-Reporting-Trials-Statement (the CONSORT-Statement) developed by JAMA, BMJ, Lancet, and Annals-of-Intern-Med in an effort to standardize the way trials are reported, with special-emphasis on the-intention-to-treat-principle in order to reduce treatment-related selection-bias. For investigators, <reporting> according to such standards will become much easier, and will even become a non-issue if requirements as requested by CONSORT are met. This initiative may have important potential to improve the level of validity of trials and thus facilitate their suitability for meta-analyses. Another important milestone is the

initiative of the Unpublished-Paper-Amnesty-Movement. In September 1997 the editors of nearly 100 international journals invited investigators to submit unpublished study data in the form of unreported-trial-registration-forms. Submitted materials are routinely made available to the world through listing the trial-details on the journals' web sites, in addition to other ways. The International-Committee-of-Medical-Editors and the World-Association-of-Medical-Editors are currently helping these initiatives by standardizing the peer review system and training referees.

Where do we go? We go for the aim of meta-analyses being accepted as gold standard for :

1. Reporting randomized experimental research.
2. Setting the stage for the development of new drugs.
3. Determination of individual therapies.
4. Leading the way for regulatory organs.
5. Maybe soon even epidemiological research.

We will only accomplish these efforts if we stick to the scientific method, which we summed up for you earlier. However, today many meta-analyses are presented or published, that do not follow these simple scientific principles, and that just leave out validity assessment of trials included, or tests for heterogeneity and publication bias. Both journal editors and readers of meta-analyses must be critical and alert since a flawed meta-analysis of unreliable and biased material is deadly, not only to research but also to health care. The above guidelines enable not only to perform meta-analyses but also to identify flawed meta-analyses, and, more importantly, to identify and appreciate well-performed meta-analyses.

8. CONCLUSIONS

The scientific methods governing the practice of meta-analysis include (1) a clearly defined prior hypothesis, (2) a thorough search of trials, (3) strict inclusion criteria, and (4) a uniform data analysis. In the statistical analysis of the meta-data three pitfalls have to be accounted: (1) publication bias, (2) heterogeneity, (3) lack of robustness.

9. REFERENCES

1. Thompson SG. Why sources of heterogeneity should be investigated. In: Chalmers I, Altman DG. Systematic reviews. BMJ Publishing Group, London, UK, 1995, pp 48-63.
2. Shipley MJ, Pocock SJ, Marmot MG. Does plasma cholesterol concentration predict mortality from coronary heart disease in elderly people? 18 year follow-up in Whitehall study. BMJ. 1991; 303: 89-92.

CHAPTER 22

META-ANALYSIS, REVIEW AND UPDATE OF METHODOLOGIES

1. INTRODUCTION

In 1982 thrombolytic therapy for acute coronary syndromes was controversial. In a meta-analysis of 7 trials Stampfer et al. found a reduced risk of mortality of 0.80 (95% confidence interval 0.68-0.95). These findings[1] were not accepted by cardiologists until 1986, when a large clinical trial confirmed the conclusions[2] , and streptokinase became widely applied.

Meta-analyses can be defined as systematic reviews with pooled data. Traditionally, they are post-hoc analyses. However, probability statements may be more valid, than they usually are with post-hoc studies, particularly if performed on outcomes that were primary outcomes in the original trials. Problems with pooling are frequent: correlations are often nonlinear[3]; effects are often multifactorial rather than unifactorial[4]; continuous data frequently have to be transformed into binary data for the purpose of comparability[5]; poor studies may be included and coverage may be limited[6] ; data may not be homogeneous and may fail to relate to hypotheses.[7] In spite of these problems, the methods of meta-analysis are an invaluable scientific activity: they establish whether scientific findings are consistent[8] , and can be generalized across populations and treatment variations[9] , and whether findings vary between subgroups.[10] The methods also limit bias, improve reliability and accuracy of conclusions[11], and increase the power and precision of treatment effects and risk exposures.[6]

The objective of this paper is to review statistical procedures for the meta-analysis of cardiovascular research. The Google data base system provides 659,000 references on the methods of meta-analysis, and refers to hundreds of books of up to 600 pages[12] , illustrating the complexity of this subject. The basic statistical analysis of meta-analyses is, however, not complex, if the basic scientific methods are met.[13] We first will review the scientific methods, and, then, introduce the statistical analysis, including the analysis of potential pitfalls. Finally, we will cover some new developments.

2. FOUR SCIENTIFIC RULES

The logic behind meta-analyses is simple and straightforward. What it requires, is to stick to scientific methods, largely similar to those required for clinical trials. They can be summarized: (1) a clearly defined prior hypothesis, (2) thorough search of trials, (3) strict inclusion criteria, and (4) uniform data analysis.[13]

Clearly defined hypothesis

Clinical trials address efficacy and safety of new drugs or interventions. It is specified in advance what are the main outcome variables, and how they should be

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 277–288.

tested. A meta-analysis is very much similar to a single trial, and, similarly to a single trial, it tests a very small number of primary hypotheses, mostly that the new compound or intervention is more efficacious and safe than the reference compound or intervention.

Thorough search of trials

The activity of thoroughly searching-published-research requires a systematic procedure, and has to be learned. You may pick up a checklist for this purpose, similarly to the checklist used by aircraft staff before take off, a nice simile used by Oxman.[14] A faulty review of trials is as perilous as a faulty aircraft and both of them are equally deadly, particularly so if we are going to use it for making decisions about health care. For a systematic review Medline[15] is not enough, and other data bases have to be searched, e.g., EMBASE-Excerpta Medica[16] and the Cochrane Library. [17]

Strict inclusion criteria

Inclusion criteria are concerned with the levels of validity, otherwise called quality criteria, of the trials to be included. Strict inclusion criteria means that we will, subsequently, only include the valid studies. Some factors have empirically been shown to beneficially influence validity. These factors include: blinding the study; random assignment of patients; explicit description of methods; accurate statistics; accurate ethics including written informed consent. We should add that the inclusion of unpublished studies may reduce the magnitude of publication bias, an issue which will be discussed in the section "Pitfalls of data-analysis".

Uniform data analysis

Statistical analysis is a tool which helps to derive meaningful conclusions from the data, and to avoid analytic errors. Statistics should be simple and test primary hypotheses in the first place. Prior to any analysis or data plots, we have to decide what kind of data we have.

3. GENERAL FRAMEWORK OF META-ANALYSIS

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

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

, and its standard error is

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

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

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

If all k studies have the same true quantitative effect, $\sigma^2 = 0$ and the weighted average effect is called a *fixed-effect* estimate. If the true effects of the compound vary between studies, $\sigma^2 > 0$ and the weighted average effect is called a *random-effects* estimate. For the fixed-effect estimate (i.e. $\sigma^2 = 0$) the calculations are quite simple, for the random-effects estimate the calculations are more complex, but available in computer packages.[18-22]

Depending on the type of outcome variable, the summary statistics x_1, x_2, ..., x_k have different forms.

1. Continuous data

Continuous data are summarized with means and standard deviations; $mean_{1i}$ and SD_{1i} in the placebo-group, and $mean_{2i}$ and SD_{2i} in the active-treatment group of trial i. The summary statistic equals $x_i = mean_{1i} - mean_{2i}$ and $se(x_i) = \sqrt{\frac{SD_{1i}^2}{n_{1i}} + \frac{SD_{2i}^2}{n_{2i}}}$, where n_{1i} and n_{2i} are the sample sizes of the two treatments.

If a trial compares two treatments in the same patients, the summary statistic is $x_i = mean_{1i} - mean_{2i}$, where $mean_{1i}$ and $mean_{2i}$ are the means of the two treatments, and $se(x_i) = \sqrt{\frac{SD_{1i}^2}{n_i} + \frac{SD_{2i}^2}{n_i} - \frac{2\, r\, SD_{1i}\, SD_{2i}}{n_i}}$, where *r* is the correlation between the outcomes in the two treatments.

If the distribution of the outcomes is very skewed, it is more useful to summarize the outcomes with medians than means.

2. Binary data

Binary data are summarized as proportions of patients with a positive outcome in the treatment arms, denoted by p_{i1} and p_{i2}. Three different summary statistics are used:

(a) Risk-difference.

The summary statistic of trial i equals $x_i=p_{1i}-p_{2i}$, the standard error equals $se(x_i)=\sqrt{\frac{p_{1i}(1-p_{1i})}{n_{1i}}+\frac{p_{2i}(1-p_{2i})}{n_{2i}}}$, where n_{1i} and n_{2i} are the sample sizes of the two treatments of trial i.

(b) Relative Risk.

The summary statistic of trial i equals the ratio of the two proportions, but its distribution is often very skewed. Therefore, we prefer to analyze the natural logarithm of the relative risk, ln(RR). The summary statistic thus equals $x_i = \ln(p_{1i} / p_{i2})$, and the standard error equals $se(x_i) = \sqrt{\frac{1-p_{1i}}{p_{1i}n_{1i}}+\frac{1-p_{2i}}{p_{2i}n_{2i}}}$.

(c) Odds Ratio.

The summary statistic of trial i equals the ratio of the odds, but since the odds ratio is strictly positive, we again prefer to analyze the natural logarithm of the odds ratio. Thus the summary statistic equals $x_i = \ln\left(\frac{p_{1i}/(1-p_{1i})}{p_{2i}/(1-p_{2i})}\right)$, and the standard error equals

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

(d) Other methods.

The Mantel-Haenszel method has been developed for the stratified analysis of odds ratios, and has been extended to the stratified analysis of risk ratios and risk differences.[23] Like the general model a weighted average effect is calculated. For the calculation of combined odds ratios Peto's method is also often used.[24] It applies a way to calculate odds ratios which may cause under- or overestimation of extreme values like odds ratios <0.2 or > 5.0.

Sometimes valuable information can be obtained from crossover studies, and, if the paired nature of the data are taken into account, such data can be included in a meta-analysis. The Cochrane Library CD-ROM provides the Generic inverse variance method for that purpose.[17]

3. Survival data

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

$\chi^2 = \left(\frac{\overline{X}_w}{se(\overline{X}_w)} \right)^2$ with one degree of freedom. A calculated χ^2-value larger than 3.841, indicates that the pooled average is significantly different from 0.0 at $p<0.05$, and, thus, that a significant different exists between the test and reference treatments. The Generic inverse variance method is also possible for the analysis of hazard ratios.[17]

4. PITFALLS OF DATA ANALYSIS

Meta-analyses will suffer from any bias that the individual studies included suffer from, including incorrect and incomplete data. Two publications underline these problems: (1) out of 49 recently published studies, 83% of the unrandomized and 25% of the randomized studies were partly refuted soon after publication[25] ; (2) out of 519 recently published trials 20% selectively reported positive results, and reported negative results incompletely.[26] Three common pitfalls of meta-analyses are listed underneath.

Publication bias

A good starting point with any statistical analysis is plotting the data (Figure 2, Chapter 21). A Christmas tree[13] or upside-down-funnel-pattern of distribution of the results of 100 published trials shows on the x-axis the mean result of each trial, on the y-axis the sample size of the trials. The smaller the trial, the wider the distribution of results. The right graph gives a simulated pattern, suggestive of publication bias: the negative trials are not published and thus missing. This cut Christmas-tree can help suspect that there is **publication bias** in the meta-analysis. Publication bias can be tested by calculating the shift of odds ratios caused by the addition of unpublished trials from abstract-reports or proceedings.[27]

Heterogeneity

In order to visually assess heterogeneity between studies several types of plots are proposed, including forest plots, radial and L' Abbe plots.[28] The forest plot of Figure 3 in Chapter 21 gives an example used by Thompson[29] of a meta-analysis with odds ratios and 95% confidence intervals (CIs), telling something about heterogeneity. On the x-axis are the results, on the y-axis the trials. We see the results of 19 trials of endoscopic intervention vs no intervention for upper intestinal bleeding: odds ratios less than one represent a beneficial effect. These trials were considerably different in patient-selection, baseline-severity-of-condition, endoscopic-techniques, management-of-bleeding-otherwise, and duration-of-follow-up. And so, this is a meta-analysis which is, clinically, very heterogeneous.

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

$$Q = \sum_{i=1}^{k} w_i (x_i - \overline{X}_w)^2 \quad \text{with k-1 degrees of freedom.}$$

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

If the above test is positive, it is common to also calculate a random-effects estimate of the weighted average, as suggested by Dersimonian and Laird.[30] We should add that, in most situations, the use of the random-effects model will lead to wider confidence intervals and a lower chance to call a difference statistically significant. A disadvantage of the random-effects analysis is that small and large studies are given almost similar weights.[31] Complementary to the Q-statistic, the amount of heterogeneity between studies is often quantified with the I^2-statistic[32]

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

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

Investigating the cause for heterogeneity

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

sample-sizes may distort the existence of association, and meta-regression techniques may be needed to investigate associations.
Outliers may also give a clue about the cause of heterogeneity. Figure 5 in Chapter 21 shows the relation between cholesterol and coronary heart disease.[33] The two outliers on top were the main cause for heterogeneity in the data.
Still other causes for heterogeneity may be involved. As an example, 33 studies of cholesterol and the risk of carcinomas showed that heterogeneity was huge.[34] When the trials were divided according to social class, the effect in the lowest class was 4 -5 times those of the middle and upper class, explaining everything about this heterogeneous result.
There is some danger of over-interpretation of heterogeneity. Heterogeneity may occur by chance, and will almost certainly be found with large meta-analyses involving many and large studies. This is particularly an important possibility when no clinical explanation is found, or when the heterogeneity is clinically irrelevant. Also, we should warn that a great deal of uniformity among the results of independently performed studies is not necessarily good; it can indicate consistency-in-bias rather than consistency-in-real-effects as suggested by Riegelman.[35]

Lack of robustness
Sensitivity or robustness of a meta-analysis is one last aspect to be addressed. When talking of strict inclusion criteria, we discussed studies with lower levels of validity. It may be worthwhile not to completely reject the studies with lower methodology.[34] They can be used for assessing sensitivity.

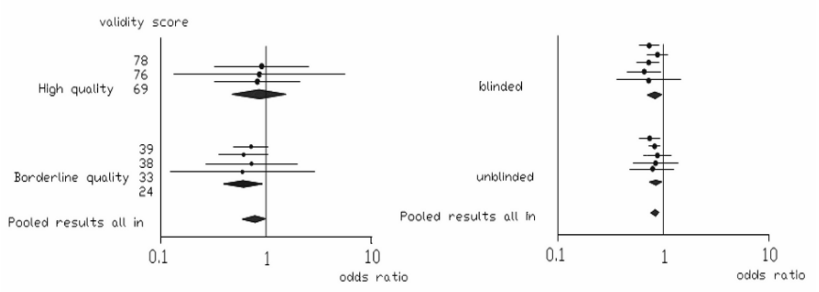


Figure 1. Left graph: three high-quality-studies provide a smaller result, than do 4 studies-of-borderline-quality; the summary result is mainly determined by the borderline-quality-studies. Right graph: when studies using objective variables are ordered according to their being blinded, differences may not be large.

The left graph of Figure 1 gives an example of how the pooled data of three high-quality-studies provide a smaller result, than do four studies-of-borderline-quality. The summary result is mainly determined by the borderline-quality-studies. When

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

5. NEW DEVELOPMENTS

Software programs for the analysis of meta-data are provided by SAS[18], the Cochrane Revman[20], S-plus[36], StatsDirect[37], StatXact[38], True Epistat.[39] Most of these programs are expensive, but common procedures are available through Microsoft's Excel and in Excel-add-ins[40] , while many websites offer online statistical analyses for free, including BUGS[41] and R[42]. Leandro's software program[43] visualizes heterogeneity directly from a computer graph based on Galbraith[44] plots.

New statistical methods are being developed. Boekholdt et al.[45] showed that observational studies and clinical trials can be simultaneously included in a meta-analysis. Van Houwelingen et al.[46] assessed heterogeneity with multivariate methods for bivariate and multivariate outcome parameters. If trials directly comparing the treatments under study are not available, indirect comparisons with a common comparator may be used.[47] A method like leave-one-out cross-validation is a standard sensitivity technique for such purpose. Lumley[48] developed network meta-analysis to compare competing treatments not directly compared in trials. Terrin et al.[49] and Tang and Liu[50] recently demonstrated that an asymmetric Christmas tree is only related to publication bias if the trials included are homogeneous, and that registries are a good alternative approach. In recent years the method of meta-regression brought new insights.[51,52] For example, it showed that group-level instead of patient-level analyses easily fails to detect heterogeneities between individual patients, otherwise called ecological biases. Robustness is hard to assess if low quality studies are lacking. Casas et al.[53] showed that it can be assessed by evaluating the extent to which different variables contribute to the variability between the studies. It can also be assessed using cumulative meta-analysis[54], while quality measures can be adjusted for in meta-regression.

Meta-analyses including few studies, e.g., 3 or 4, have little power to test the pitfalls. In contrast, meta-analyses including many studies may have so much power that they demonstrate small pitfalls, that are not clinically relevant. For example, a meta-analysis of 43 angiotensin blocker studies[55] found 95% confidence intervals of the heterogeneity and publication bias effects were not

wider than 5% of the treatment effects. Another reason why the pitfalls receive less attention today than 5 years ago is, that an increasing part of the current meta-analyses are performed in the form of working papers of an explorative nature, where the primary question is not a result representative for the entire population, but rather the estimates of the treatment effects in subgroups and interactions. These meta-analyses contain many details, and look a bit like working papers of technological evaluations as produced by physicists. The trend to increasingly publish detailed data, rather than study reports as allowed by journals, is enhanced by the Internet, which enables to register many more data than do medical journals. Meta-analyses were 'invented' in the early 70s by psychologists, but pooling study results extends back to the early 1900s by statisticians such as Karl Pearson, and Ronald Fisher. In the first years pooling of the data was often impossible due to heterogeneity of the studies. However, after 1995 trials became more homogeneous. In the late 90s several publications concluded that meta-analyses did not accurately predict treatment [56,57] and adverse effects.[58] The pitfalls were held responsible. Initiatives against them include (1) the Consolidated-Standards-of-Reporting-Trials-Movement (CONSORT), (2) the Unpublished-Paper-Amnesty-Movement of the English journals, and (3) the World Association of Medical Editors' initiative to standardize the peer review system. Guidelines / checklists for reporting meta-analyses were published like QUOROM (Quality of Reporting of Meta-analyses) and MOOSE (Meta-analysis Of Observational Studies in Epidemiology).

6. CONCLUSIONS

Meta-analysis is important in cardiovascular research, because it establishes whether scientific findings are consistent, and can be generalized across populations. The statistical analysis consists of the computation of weighted averages of study characteristics and their standard errors. Common pitfalls of data-analysis are (1) publication bias, (2) heterogeneity, (3) lack of robustness. New developments in the statistical analysis include (1) new software easy to use, (2) new arithmetical methods that facilitate the assessment of heterogeneity and comparability of studies, (3) a current trend towards more extensive data reporting including multiple subgroup and interaction analyses. Meta-analyses are governed by the traditional rules for scientific research, and the pitfalls are, particularly, relevant to hypothesis-driven meta-analyses, but less so to current working papers with emphasis on entire data coverage.

7. REFERENCES

1. Stampfer MJ, Goldhaber SZ, Yusuf S. Effects of intravenous streptokinase on acute myocardial infarction: pooled results from randomized trials. N Engl J Med. 1982; 307:1180-2.
2. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute

myocardial infarction. Lancet. 1986; 1: 397-402.
3. Glass GV, Smith ML. Meta-analysis of research on class size and achievement. Educational Evolution and Policy Analysis. 1979; 1: 2-16.
4. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer. J Clin Epidemiol. 1991; 44: 127-9.
5. Stein RA. Meta-analysis from one FOA reviewer's perspective. Proceedings of the Biopharmaceutical Section of the American Statistical Association. 1988; 2: 34-8.
6. Zhou X, Fang J, Yu C, Xu Z, Lu Y. Meta-analysis. In: Advanced Medical Statistics, Lu Y and Fang J, eds, World Scientific, River Edge, NJ, 2003, pp 233-316.
7. Turnbul F for the Blood Pressure Lowering Trialists' Collaboration. Effects of different blood pressure lowering regimens on major cardiovascular events: results of prospectively designed overviews of randomised controlled trials. Lancet. 2003; 362: 1527-35.
8. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of the best evidence for clinical decisions. Ann Intern Med. 1998; 317: 339-42.
9. Straus SE, Sackett DL. Using research findings in clinical practice. BMJ. 1998; 317: 339-42.
10. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. BMJ. 1998; 317: 465-8.
11. Jones DR. Meta-analysis: Weighing the evidence. Stat Med. 1995; 14: 137-9.
12. Hunter JE, Schmidt FL. Methods in meta-analysis. 2nd Edition. Sage Public Inc, NY, 2004.
13. Cleophas TJ. Meta-analysis. In: Cleophas TJ, Zwinderman AH, Cleophas AF, eds, Statistics Applied to Clinical Trials, Third Edition, Springer, NY, 2006, pp 205-18.
14. Oxman AD, Guyatt G. Guidelines for reading reviews. Can Med Assoc J. 1988; 138: 697-703.
15. Greenhalgh T. How to read a paper The Medline database. BMJ.1997; 315: 180-3.
16. Lefebvre C, McDonald S. Development of a sensitive search strategy for reports of randomized trials in EMBASE. In: Paper presented at the Fourth International Cochrane Colloquium, 20-24 Oct 1996; Adelaide, Australia, 1996.
17. Cochrane Library. http://www.cochrane.org/cochrane/hbook.htm
18. SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas
19. SPSS Statistical Software. http://www.spss.com
20. Cochrane Revman. http://www.cochrane.org/cochrane/revman.htm
21. Stata, statistical software for professionals. http://www.stat.com
22. Comprehensive Meta-analysis, by Biostat. http://www.meta-analysis.com
23. Greenland S, Robins JM. Estimation of common effect parameter from sparse

follow-up data. Biometrics. 1985; 41: 55-68.
24. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis. 1985; 27: 335-71.
25. Ioannides JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005; 294: 210-28.
26. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on Pub Med : review of publications and survey of authors. BMJ. 2005; 330: 753-6.
27. Chalmers I, Altman DG. Systematic reviews. BMJ Publishing Group, London, UK, 1995.
28. National Council of Social Studies. Statistical and power analysis software. http://www.ncss.com/metaanal.html
29. Thompson SG. Why sources of heterogeneity should be investigated. In: Chalmers I, Altman DG. Systematic reviews. BMJ Publishing Group, London, UK, 1995, pp 48-63.
30. Dersimonian R, Laird NM. Meta-analysis in clinical trials. Control Clin Trials. 1986; 7: 177-88.
31. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. Stat Med.1989; 8: 141 -51.
32. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21: 1539-58.
33. Shipley MJ, Pocock SJ, Marmot MG. Does plasma cholesterol concentration predict mortality from coronary heart disease in elderly people? 18 year follow-up in Whitehall study. BMJ. 1991; 303: 89-92.
34. Khan KS. Daya S, Jadad AR. The importance of quality of primary studies in producing unbiased systematic reviews. Arch Int Med. 1996; 156: 661-6.
35. Riegelman RK. Meta-analysis. In: Studying a study & testing a test. Riegelman RK, ed, Lippincott Williams & Wilkins, Philadelphia, PA, USA, 2005, pp 99-115.
36. S-plus.http://www.mathsoft.com/splus
37. StatsDirect. http://www.camcode.com
38. StatXact. http://www.cytel.com/products/statxact/statact1.html
39. True Epistat. http://ic.net/~biomware/biohp2te.htm
40. Meta-analysis Mark X. Microsoft's Excel
41. BUGS y WinBUGS. http://www.mrc-bsu.cam.ac.uk/bugs
42. R. http://cran.r-project.org
43. Leandro G. Meta-analysis in medical research. BMJ books, London UK, 2005.
44. Galbraith RF. A note on graphical presentation of estimated odds ratios from several trials. Stat Med. 1988; 7: 889-94.
45. Boekholdt SM, Sacks FM, Jukema JW, Shepherd J, Freeman DJ, McMahon AD, Cambien F, Nicaud V, De grooth GJ, Talmud PJ, Humphries SE, Miller GJ, Eiriksdottir G, Gudnason V, Kauma H, Kakko S, Savolainen MJ, Arca M, Montasli A, Liu S, Lanz HJ, Zwinderman AH, Kuivenhoven JA, Kastelein JJ.

Cholesterol ester transfer protein TaqiB variant, high density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment. Circulation. 2005; 111: 278-87.
46. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta -analysis: multivariate approach and meta-regression. Stat Med. 2002; 21: 589 -624.
47. Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico RD, Bradburn M, Eastwood AJ. Indirect comparisons of competing interventions. Health Technol Assess. 2005; 9: 1-148.
48. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002; 21: 2313-24.
49. Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP. External validation of predictive models: a comparison of logistic regression, classification trees, and neural networks. J Clin Epidemiol. 2003; 56: 721-9.
50. Tang JL, Liu JL. Misleading funnel plots for detection of bias in meta-analysis. J Clin Epidemiol. 2000; 53: 477-84.
51. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol. 2004; 57: 683-97.
52. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med. 2004; 23: 1662-82.
53. Casas JP, Leonelo EB, Humphries SE. Endothelial NO synthase genotype and ischemic heart disease. Circulation. 2004; 109: 1359-65.
54. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol. 1995; 48: 45-57.
55. Conlin PR, Spence JD, Williams B, Ribeiro AB, Saito I, Benedict C, Bunt AM. Angiotensin II antagonists for hypertension: are there differences in efficacy? Am J Hypertens. 2000; 13: 418-26.
56. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderien F. Discrepancies between meta-analyses nand subsequent large randomized controlled trials. N Engl J Med. 1997; 337: 536-42.
57. Temple R. Meta-analyses and epidemiological studies in drug development and postmarketing studies. JAMA. 1999; 281: 841-4.
58. Brewer T, Colditz GA. Postmarketing surveillance and adverse drug reactions; current perspectives. JAMA. 1999; 281: 824-9.

CHAPTER 23

CROSSOVER STUDIES WITH CONTINUOUS VARIABLES

1. INTRODUCTION

Crossover studies with continuous variables are routinely used in clinical drug research: for example, no less than 22% of the double-blind placebo-controlled hypertension trials in 1993 were accordingly designed.[1] A major advantage of the crossover design is that it eliminates between-subject variability of symptoms. However, problems include the occurrence of carryover effect, sometimes wrongly called treatment-by-period interaction (see also chapter 17): if the effect of the first period carries on into the next one, then it may influence the response to the latter period. Second, the possibility of time effects due to external factors such as the change of the seasons has to be taken into account in lengthy crossover studies. Third, negative correlations between drug responses, although recently recognized in clinical pharmacology, is an important possibility not considered in the design and analysis of clinical trials so far. Many crossover studies may have a positive correlation-between-drug-response, not only because treatments in a given comparison are frequently from the same class of drugs, but also because one subject is used for comparisons of two treatments. Still, in treatment comparisons of completely different treatments patients may fall into different populations, those who respond better to the test-treatment and those who do so to the reference-treatment. This phenomenon has already lead to treatment protocols based on individualized rather than stepped care.[2] Power analyses for crossover studies with continuous variables so far only accounted for the possibility of approximately zero levels of correlations.[3-8] While considering different levels of correlation, we recently demonstrated[9] that the crossover design with binary variables is a powerful means of determining the efficacy of new drugs in spite of such factors as carryover effects. Crossover trials with continuous variables, however, have not yet been similarly studied.

In the current communication while taking both positive and negative correlations into account we drew power curves of hypothesized crossover studies with different amounts of treatment effect, carryover effect and time effect.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 289–298.

2. MATHEMATICAL MODEL

According to Scheffé[10] the notion for a simple two-period two-group crossover study is

	period 1		period 2	
	treatment	mean effect	treatment	mean effect
Group 1 (n_1)	1	$y_{1.1}$	2	$y_{1.2}$
Group 2 (n_2)	2	$y_{2.1}$	1	$y_{2..2}$

where y_{ijk} = the response in the jth patient in the ith group in the kth period. We assume that $n_1 = n_2 = n$ and that we have normal distributions or t-distributions. $y_{i.k} = \sum y_{ijk} / n$.

Treatment, carryover and time effects are assessed according to Grizzle.[11] To test treatment effect φ the sum of the results of treatment 1 is compared with the treatment 2 results ($y_{1.1} + y_{2.2}$ versus $y_{1.2} + y_{2.1}$). To trace carryover effect (λ) the sum of the results in group 1 is compared with the group 2 results ($y_{1.1} + y_{1.2}$ versus $y_{2.1} + y_{2.2}$). To trace time effect (π) the sum of the results in period 1 is compared with the period 2 results ($y_{1.1} + y_{2.1}$ versus $y_{1.2} + y_{2.2}$).

The null-hypotheses that φ, λ, and π are zero

$$\varphi \quad [(y_{1.1} + y_{2.2}) - (y_{1.2} + y_{2.1})] = 0$$
$$\lambda \quad [(y_{2.1} + y_{2.2}) - (y_{1.1} + y_{1.2})] = 0$$
$$\pi \quad [(y_{1.1} + y_{2.1}) - (y_{1.2} + y_{2.2})] = 0$$

should be slightly remodeled into paired comparisons, because otherwise calculations cannot be appropriately accomplished.

$$\varphi \quad [(y_{1.1} - y_{1.2}) - (y_{2.1} - y_{2.2})] = 0$$
$$\lambda \quad [(y_{2.1} + y_{2.2}) - (y_{1.1} + y_{1.2})] = 0$$
$$\pi \quad [(y_{1.1} - y_{1.2}) + (y_{2.1} - y_{2.2})] = 0$$

In this way 2 x 2 paired cells can be adequately added or subtracted in a cell by cell manner.

3. HYPOTHESIS TESTING

These null hypotheses can be tested, for example, by paired t-statistic or repeated measures analysis of variance (ANOVA). The larger the extent to which the t or F value of our distribution differs from zero, the more sensitivity the statistical approach does provide.

$$t = \frac{d}{SE} \quad \text{(or repeated measures ANOVA, F value)}$$

where d is φ, λ, or π, and SE is their standard error.
SE is calculated by use of the standard formulas for the variance ($\sqrt{\sigma^2}/n$) of paired and unpaired sums and differences.

$$\sigma^2_{\text{paired sums}} = \sigma_1^2 + \sigma_2^2 + 2\rho\,\sigma_1\,\sigma_2$$
$$\sigma^2_{\text{paired differences}} = \sigma_1^2 + \sigma_2^2 - 2\rho\,\sigma_1\,\sigma_2$$
$$\sigma^2_{\text{unpaired sums}} = \sigma_1^2 + \sigma_2^2$$
$$\sigma^2_{\text{unpaired differences}} = \sigma_1^2 + \sigma_2^2$$

If we assume that $\sigma = \sigma_{Y1.1} = \sigma_{Y1.2} = \sigma_{Y2.1} = \sigma_{Y2.2}$ = standard deviation of the samples in each of the cells, and that $\rho = \rho_{Y1.1 \text{ vs } Y1.2} = \rho_{Y2.1 \text{ vs } Y2.2}$ = correlation coefficient between the samples of each of the two paired cells, then

$$\sigma_\varphi^2 = 2(2\,\sigma^2)(1-\rho)$$
$$\sigma_\lambda^2 = 2(2\,\sigma^2)(1+\rho)$$
$$\sigma_\pi^2 = 2(2\,\sigma^2)(1-\rho)$$

Because $n_1 = n_2 = n$, we now can calculate the SEs as follows:

$$SE_{\varphi} = \sqrt{4\sigma^2(1-p)(\frac{1}{2n}+\frac{1}{2n})} = \sqrt{\frac{4\sigma^2(1-p)}{n}}$$

and accordingly

$$SE_{\lambda} = \sqrt{\frac{4\sigma^2(1+p)}{n}}$$
$$SE_{\pi} = \sqrt{\frac{4\sigma^2(1-p)}{n}}$$

Suppose $\lambda=\varphi$ and $\rho= 0$, then $t_{\lambda} = t_{\varphi\_}$. In this situation the sensitivity to test carryover and treatment effect are equal.

If $\lambda = \varphi$ and $\rho > 0$ then $t_{\lambda} < t_{\varphi\_}$
If $\lambda = \varphi$ and $\rho < 0$ then $t_{\lambda} > t_{\varphi}$

So, the sensitivity of testing is largely dependent on the correlation between treatment modalities ρ. Whenever $\rho > 0$ we soon will have a much larger t-value, and, thus, better sensitivity to test treatment effect than carryover effect of similar size. We should add that in practice $\sigma_{Y1.2}$ may be somewhat larger than $\sigma_{Y1.1}$, because the larger the data the larger the variances. If, e.g., $\sigma_{Y1.2}$ is 10% larger than $\sigma_{Y1.1}$, ρ will change from 0.00 to 0.05. So, in this situation the level of positive correlation required tends to rise.
Time effect (π) is generally considered to influence one treatment similarly to the other, and its influence on the size of the treatment difference is, thus, negligible.

	Period 1		Period 2	
	Treatment	Mean response	Treatment	Mean response
Group 1	1	$y_{1.1}$	2	$y_{1.2} + ½\pi$
Group 2	2	$y_{2.1}$	1	$y_{2.2} + ½\pi$

Under the assumption $\varphi = 0$ we have
$\varphi = (y_{1.1} - y_{1.2} - ½\pi) - (y_{2.1} - y_{2.2} - ½\pi)$
$= y_{1.1} - y_{1.2} - y_{2.1} + y_{2.2}$

Although time or period effects may introduce extra variance in the study, the crossover design in a way adjusts for time effects, and some even believe that time effects do not have to be taken into account in the routine analysis of crossover studies, unless there is a clinical interest to know.[7]

4. STATISTICAL POWER OF TESTING

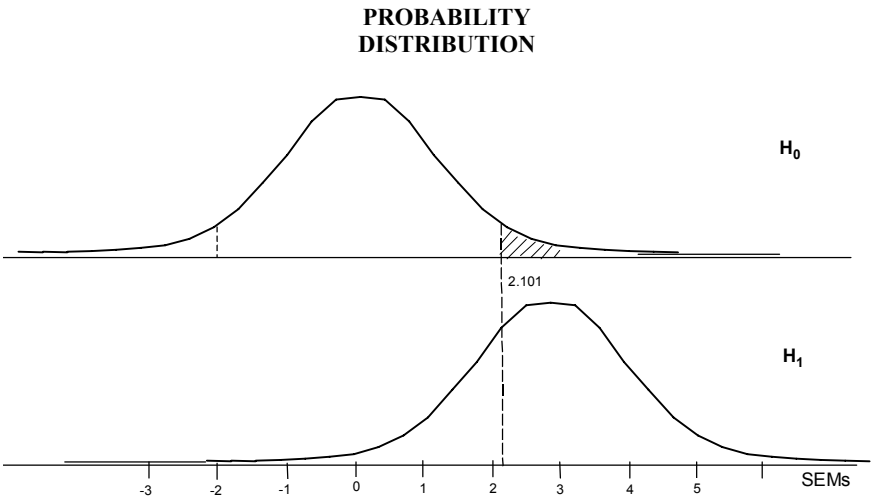


Figure 1. Example of a t-distribution (H_1) and its null hypothesis (H_0). α = % chance of erroneously rejecting this null hypothesis (usually taken as 5%), β= % chance of erroneously accepting this null hypothesis. Statistical power is defined as (1-β) x 100%.

Figure 1 gives an example of a t-distribution (H_1) and its null hypothesis of no effect (H_0). α = % chance of erroneously rejecting this null hypothesis (usually taken as 5%), and β=% chance of erroneously accepting this null hypothesis. Statistical power is defined as (1-β) x 100%. Statistical power can be approximated from the equation (prob = probability):

$$\text{POWER} = 1 - \beta = 1 - \text{prob}\,[Z \leq (t - t^1)]$$

where Z represents the standardized value for the differences between mean and zero and t^1 represents the upper critical value of t for the given degrees of freedom and α has been specified ($\alpha = 0.05$).
Suppose we have a crossover study with n=10 per group, because this is a size frequently used in such studies, and with $\varphi = \sigma$ = standard deviation of the samples in each cell, because this is frequently approximately so. Then increasing amounts of λ are added with $\sigma_\lambda = \lambda$. The influence of this procedure on the statistical power of testing λ and φ are then assessed. The amounts of λ are expressed as λ/φ ratios. Power graphs are calculated for three different levels of correlation-between-drug-response ($\rho \cong -1$; $\rho \cong 0$; $\rho \cong +1$).

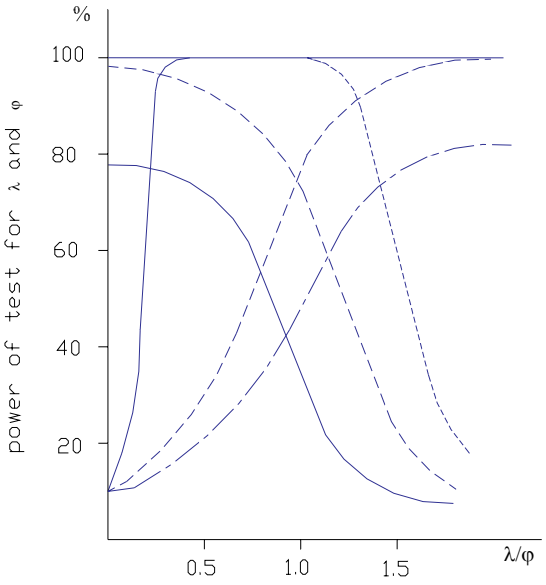


Figure 2. Statistical power of testing carryover effect (slope upwards) and treatment effect (slope downwards); λ = carryover effect, φ = treatment effect, ρ = correlation coefficient.

_____ $\rho \cong -1$
_ _ _ _ _ _ _ _ $\rho \cong 0$
---------- $\rho \cong +1$

Figure 2 shows the results. First, there are three power curves of treatment effect for the three levels of correlation. As λ/φ increases, all three gradually come down. The negative correlation curve is the first to do so. Consequently, this situation has generally little power of rightly coming to the right conclusion. At λ/φ= 1.0, when treatment effect is equal to carryover effect, there is less than 30% power left. It means we have a more than 70% chance that treatment effect is erroneously unobserved in this study. Considering that a power of approximately 80% is required for reliable testing, we cannot test carryover here in a sensitive manner. The zero and positive correlation situations provide essentially better power.

There are also three power curves of carryover effect for three correlation levels. The negative correlation curve provides essentially better power than the zero and positive correlation curves do. This example shows that strong positive correlations leave little power to test carryover effect. It also shows that strong negative correlations produce excessive power to test carryover effect.

The amounts of time effect are generally assumed to influence the two treatment groups similarly, and it, therefore, may hardly influence the treat comparison. Suppose in the above example time effect (π) instead of carryover effect (λ) is added in increasing amounts with $\sigma_\pi = \pi$.

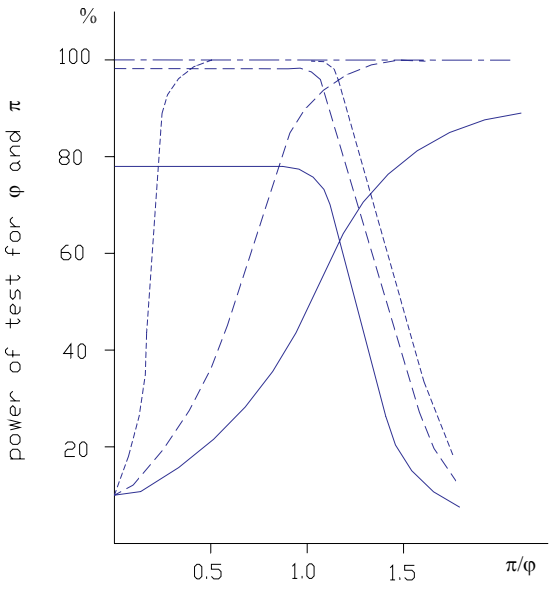


Figure 3. Statistical power of testing time effect (slope upwards) and treatment effect (slope downwards). π = time effect, φ = treatment effect, ρ = correlation coefficient.

———————— $\rho \cong -1$
— — — — — — — — $\rho \cong 0$
---------- $\rho \cong +1$

Figure 3 shows the influence of increasing ratios π / φ on the statistical power of testing π and φ. First, small time effects unlike carryover effects hardly influence nor the amount nor the statistical power of testing treatment effect. Also the power of demonstrating time effect is largely dependent on the level of correlation-between-drug-response: with a negative correlation we have little power to demonstrate time-effect. In contrast, with a positive correlation we have a lot of power to do so.

We conclude that the level of correlation-between-drug-response is a major determinant of not only the power of demonstrating treatment effect but also that of time effect in the current approach.

5. DISCUSSION

The crossover design for treatment comparisons with continuous variables provides approximately equal statistical power to test carryover, time, and treatment effects when between-treatment correlation is not strong positive/negative. E.g., in the hypothesized crossover situation from our example the statistical power to demonstrate similarly-sized treatment and carryover, or treatment and time effects is approximately 80% (as demonstrated in the above figures), which is generally considered to be an acceptable level for reliable testing. However, whenever the correlation coefficient is >0, we will soon have better sensitivity to test treatment than carryover or time effect of similar size. Inversely, whenever it is <0, we will soon have better sensitivity to demonstrate the latter two rather than the former.
We should add that calculations are made under the assumption that either carryover or time effect are in the study. If both effects are simultaneously in the study, variances have to be added up and powers will be somewhat smaller. The assumption does not invalidate the overall conclusion of the procedure as it produces the largest powers for the given data.

Analysis of covariance (ANCOVA)

Analysis of covariance is used if two x-variables are dependent on one another.
When F-tests are used instead of t-tests, the sensitivity of testing can be somewhat improved by analysis of covariance (ANCOVA) according to

$$\text{adjusted } SS_{\text{treatment}} \text{ between groups} = \text{unadjusted } SS_{\text{treatment}} \text{ between groups} + (\text{SP within groups})^2 / SS_{\text{carryover}} \text{ within groups} - (\text{SP total})^2 / SS_{\text{carryover}} \text{ total}$$

$$\text{adjusted SS within groups} = \text{unadjusted SS within groups} - (\text{SP within groups})^2 / SS_{\text{carryover}} \text{ within groups}$$

where SS = sum of squares, and SP = sum of products of treatment by carryover effects (treatment effect x carryover effect).

Computation can be found, e.g., in Hays' textbook Statistics[12], and can be readily made by statistical packages, e.g., SPSS[13] under the subprogram "ANOVA".
In this way, power of testing may improve by a few percentages. However, this method of adjustment can be used only when correlations are not strong + or - , and when n is at least 20 or more, which is not so in many crossover studies. Also the method only adjusts statistical sensitivity, but not amounts of treatment, carryover or time effects, and so its usefulness is limited.
Although the analysis uses multiple comparisons testing, the p-values do not have to be multiplied by the number of tests, because although the chance of a positive

test increases, the chance of e.g., a positive test for carryover does not as it is only tested once.

The current chapter stresses the major impact of correlation level between treatment comparison, and particularly the phenomenon of negative correlations. This phenomenon is only shortly being recognized and may have flawed many trials so far. In a trial the test treatment is frequently a slight modification of the reference treatment or is equivalent to it with addition of just a new component. In this situation there is obviously a positive correlation between responses to test and reference treatments. However, completely new classes of drugs are continually being developed and are tested against established classes of drugs. With the comparison of drugs from completely different classes patients may fall into different populations: those who respond better to one class and those who do so to the other class.. E.g., patients with angina pectoris unresponsive to calcium channel blockers or nitrates. may respond very well to beta blockers. Also hypertension, cardiac arrhythmias, chronic obstructive pulmonary disease are conditions where a non-response is frequently associated with an excellent response to a completely different compound. These are situations where a crossover study may give rise to a strong negative correlation. It would mean that a crossover design for the comparisons of treatment from completely different classes of drugs is endangered of being flawed and that such comparisons had better be assessed in the form of a parallel group comparison which evens out within subject variability.

6. CONCLUSION

Background: The crossover design is a sensitive means of determining the efficacy of new drugs because it eliminates between subject-variability. However, when the response in the first period carries on into the second (carryover effects) or when time factors can not be kept constant in a lengthy crossover (time effects), the statistical power of testing may be jeopardized. We recently demonstrated that the crossover design with binary variables is a powerful method in spite of such factors as carryover effects.[9] Power analysis of crossover trials with continuous variables has not been widely published.
Objective: Using the Grizzle model for the assessment of treatment effect, carryover effect and time effect, we drew power curves of hypothesized crossover studies with different levels of correlation between drug responses.
Results: We demonstrate that the sensitivity of testing is largely dependent on the levels of correlation between drug response. Whenever the correlation coefficient is >0, we soon will have better sensitivity to test treatment effect than carryover effect or time effect of similar size. Whenever levels of correlation are not strong positive or negative the statistical power to demonstrate similarly-sized treatment and carryover effect, or treatment and time effect is approximately 80%, which is an acceptable level for reliable testing.
Conclusions: The crossover design is a powerful method for assessing positively correlated treatment comparisons, despite the risk of carryover and time effects.

7. REFERENCES

1. Niemeyer MG, Zwinderman AH, Cleophas TJ, De Vogel EM. Crossover studies are a better format for comparing equivalent treatments than parallel-group studies. In: Kuhlmann J, Mrozikiewicz A, eds, What should a clinical pharmacologist know to start a clinical trial (phase I and II). Munich, Germany, Zuckschwerdt Verlag, 1998, pp 40-48.
2. Scheffé H. Mixed models. In: Scheffé H, ed, The analysis of variance. New York, Wiley & Sons, 1959, pp 261-91.
3. Cleophas TJ. Crossover studies: a modified analysis with more power. Clin Pharmacol Ther 1993; 53: 515-20.
4. Willan AR,Pater JL. Carryover and the two-period crossover clinical trial . Biometrics 1986; 42: 593-9.
5. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials . Stat Med 1989; 8: 1421-32.
6. Fleiss JA. A critique of recent research on the two-treatment crossover design . Control Clin Trials 1989; 10: 237-41.
7. Senn S. The AB/BA crossover: past, present and future. Stat Methods Med Res 1994; 3: 303-24.
8. Grieve AP. Bayesian analyses of two-treatment crossover studies. Stat Methods Med Res 1994; 3: 407-29.
9. Cleophas TJ, Van Lier HH. Clinical trials with binary responses: power analyses. J Clin Pharmacol 1996; 36: 198-204.
10. Nies AS, Spielberg SP. Individualization of drug therapy. In: Hardman JL et al., eds, Goodman and Gilman's Pharmacological Basis of Therapeutics . New York: McGraw-Hill,1996, pp 43-63.
11. Grizzle JE. The two-period change-over design and its use in clinical trials. Biometrics 1965; 22: 469-80.
12. SPSS 8 for Windows 95 and 98, SPSS Benelux, Gorinchem, Netherlands.
13. Hays WL. Statistics. Fort Worth, TX, Holt, Rinehart and Winston, Inc, 4th edition, 1988.

CHAPTER 24

CROSSOVER STUDIES WITH BINARY RESPONSES

1. INTRODUCTION

The crossover design is widely used in clinical research especially in the case of a limited number of patients. The main advantage of within-patient over between-patient comparisons is that between-subject variability is not used in the comparisons. However, a prerequisite is that the order of the treatments does not influence the outcome of the treatment. If the effect of the treatment administered in the 1st period carries on into the 2nd period, then it may influence the measured response in the 2nd period. This essentially means that only symptomatic treatments qualify for crossover comparisons and curative treatments do not. However, symptomatic treatments frequently have small curative effects, e.g., wound healing by vasodilators or, more recently, cardiac remodelling by after load reduction. The treatment group that is treated with the effective compound first and with the less effective compound or placebo second is frequently biased by carryover effect from the 1st period into the 2nd, whereas the alternative group that is treated in the reverse order is not so.[1] For example, of 73 recently published crossovers only 6 reported the data of the separate periods. In 5 of them (83%) this very type of carryover effect was demonstrable. Such a mechanism may cause a severe underestimation of the treatment results[2] and this possibility should, therefore, be assessed in the analysis. Most of the reports on the subject of order effects so far have addressed crossover studies with a quantitative rather than binary response. [3-10] Although Hills & Armitage [11] in an overview of methods in crossover clinical trials mentioned the tests of Gart [12] and Prescott[13] for crossover trials with a binary response and Fidler [14] presented a model, little attention has been paid to this kind of trials. A binary response is different from a quantitative in that it generally does not answer what exactly can be expected in an individual. Rather it addresses whether or not a particular result has a predictive value, which one of two treatments is better, or whether there is a treatment effect in the data. One might contend, therefore, that some undervaluation of a difference in binary data is not that important as long as it does not cause a type II error of finding no difference were there is one. The main issue of the present paper is the question whether in a crossover trial with a binary response a significant carryover effect does leave enough power in the data to demonstrate a treatment effect.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 299–307.

2. ASSESSMENT OF CARRYOVER AND TREATMENT EFFECT

In a crossover trial with two treatments and two periods the patients are randomized into two symmetric groups that are treated with treatments A and B in a different order (table 1). If groups are symmetric and the results are not influenced by the order of the treatments, the probabilities of treatment success in group I and II should be virtually the same in each period for each treatment: p_A being the probability of treatment success from treatment A, p_B from treatment B (Table 1).

Table 1. Example of a crossover design with a binary response

	Period 1		Period II	
	Treatment	**Probability of treatment success**	**Treatment**	**Probability of treatment success**
Group I	A	p_A	B	p_B
Group II	B	p_B	A	p_A*

* If in Group II treatment B has a carryover effect on the outcome of treatment A, p_A changes to p_C. If $P_B = p_C$, carryover effect is maximal.

The group that is treated with the less effective treatment or placebo after the more effective is endangered of being biased by carryover effect from the 1st period into the 2nd.

Suppose treatment A is far less effective than B (table 1). Then, if in Group II treatment B has a carryover effect on the outcome of treatment A, the probability of treatment success changes from p_B into p_C. To detect a carryover effect we compare the outcomes of treatment A in Group I to those in group II: p_A versus p_C, an unpaired comparison. The amount of carryover effect in group II is considered to be the difference between p_C and p_A. Carryover effect in Group I (ineffective treatment period prior to effective) is assumed to be negligible. Time effect is assumed to be negligible as well, because we study stable disease only. It thus seems that neither a test for carryover effect in Group I, nor a test for time effects needs to be included in our assessment. Treatment effect is assessed by taking the two groups together after which all of the outcomes of the treatments A are compared with those of the treatments B in a paired comparison. The assumption that carryover effect is negligible implies that the test for carryover effect uses only half of the available data and might therefore be expected to be less sensitive. However, sensitivity not only depends on sample size but also on the size of differences and their variances.

3. STATISTICAL MODEL FOR TESTING TREATMENT AND CARRYOVER EFFECTS

We assume an unidirectional assessment where p is between 0.0 (no symptoms anymore) and 1.0 (=100% remains symptomatic in spite of treatment). When carryover effect is in the data, p_A in Group II turns into p_C (table 1). The difference between p_C and p_A is considered to be the amount of carryover effect in the data. Fisher exact test, as explained in chapter 3, is used for testing whether p_C is significantly different from p_A. With the program of Bavry [15] those values of p_C are determined that should yield a significant carryover effect in 80% of the trials (i.e. the power equals 80%). The number of patients in both groups is chosen between 10 and 25, because many crossover trials have 20 to 50 patients. These values of p_C are then used for determining whether in crossover trials with significant carryover effect and a binary response enough power is left in the data for demonstrating a significant treatment effect.

For testing the treatment effect all of the data of the treatment A are taken together and compared with those of the treatments B. The power of this test depends not only on the probabilities p_A and p_B, but also on the correlation between the treatment responses. This correlation is expressed as $\rho = p_{A/B} - p_A$, where $p_{A/B}$ is the probability of a treatment success with A, given that treatment B was successful. When $\rho = 0$, treatments A and B act independently. When p_B equals p_C, this would mean that carryover effect in group II is not only significant but also maximal given the amount of treatment effect. Considering this situation of maximal carryover effect, we calculate the power of detecting treatment effects. The power of McNemar's test with p_B being equal to p_C and with various values of p_A was calculated according to Bavry [15] .

4. RESULTS

CALCULATION OF PC VALUES JUST YIELDING A SIGNIFICANT TEST FOR CARRYOVER EFFECT

For various numbers of patients and various values of p_A (the probability of success with treatment A in period I, Table 1), the p_C values (the probability of success with treatment A in period II) are calculated that with a power of 80% will give a significant test for carryover effect (p_A versus p_C, $\alpha = 0.05$).

Table 2 shows that carryover effects (difference between p_A and p_C) as large as 0.60, 0.50, 0.40 and 0.35 are required for a significant test. For $\alpha = 0.01$, these values are about 0.70, 0.60, 0.50 and 0.45. Using these p_C values, we then calculated the probability of detecting a treatment effect (i.e. power of testing treatment effect). We report minimal values of power only, i.e., the situation where $p_B = p_C$. Whenever $p_B < p_C$, we would have even better power of testing treatment effect.

Table 2. Power to demonstrate a treatment effect in spite of the presence of a significant carryover effect

	Total number of patients			
p_A	**2 x 10**	**2 x 15**	**2 x 20**	**2 x 25**
0.10				
0.20				
0.30				98 (0.02)
0.40		96 (0.02)	97 (0.05)	96 (0.08)
0.50		97 (0.06)	96 (0.11)	96 (0.14)
0.60	97* (0.04)#	98 (0.11)	96 (0.18)	95 (0.23)
0.70	96 (0.11)	97 (0.20)	97 (0.26)	94 (0.33)
0.80	96 (0.20)	97 (0.30)	97 (0.37)	96 (0.43)
0.90	96 (0.31)	97 (0.43)	96 (0.47)	96 (0.52)

* Power (%) of McNemar's test for treatment effect ($\alpha = 0.05$, $\rho = 0$).
\# p_C value just yielding a significant test for carryover effect ($\alpha = 0.05$, power = 80%).

POWER OF PAIRED COMPARISON FOR TREATMENT EFFECT

When the result of treatment B (p_B) is taken equal to the maximal values of p_C and treatments A and B act independently ($\rho = 0$), the probability of detecting a treatment effect (i.e. the power) in the crossover situation with n between 20 and 50 is always more than 94% (Table 2). Usually, however, treatments A and B do not

act independently. With a negative correlation between the two treatments modalities power is lost, with a positive correlation it is augmented. Table 3 shows power values adjusted for different levels of ρ . With negative levels of ρ and 20 patients the power for detecting a treatment difference is not less than 74% which is about as large as that chosen for the test on carryover effect (80%). When more patients are admitted to the trial this value will be about 90%.

Table 3. Power (%) to demonstrate a treatment effect in spite of the presence of a significant carryover effect

		Total number of patients			
	ρ	**2 x 10**	**2 x 15**	**2 x 20**	**2 x 25**
α_1^* = 0.05	-0.20	89	94	96	95
α_2 = 0.05	-0.10	92	96	97	97
	0	96	96	96	94
	0.10	98	97	98	99
	0.20	98	98	99	99
α_1 = 0.01	-0.20	95	99	94	99
α_2 = 0.01	-0.10	97	100	99	99
	0	99	99	99	99
	0.10	100	100	100	100
	0.20	100	100	100	100
α_1 = 0.10	-0.20	74	84	89	88
α_2 = 0.05	-0.10	79	91	92	90
	0	85	90	89	88
	0.10	89	95	95	94
	0.20	95	94	97	97
α_1 = 0.05	-0.20	75	87	90	90
α_2 = 0.01	-0.10	81	92	92	93
	0	88	90	90	89
	0.10	92	93	95	96
	0.20	96	96	98	98

* α_1 level of significance of test for carryover effect.
α_2 level of significance of test for treatment effect.
ρ level of correlation between treatments A and B.

5. EXAMPLES

Suppose we have a negative crossover where probability of treatment success group II p_C (Table 4) may have changed from 0.8 into 0.2 due to carryover effect from the effective treatment B into the 2nd period. Fisher exact test for demonstrating a carryover effect (p_A versus p_C) is calculated according to

$$\text{Point probability for carryover effect} = \frac{10!\ 10!\ 10!\ 10!}{20!\ 2!\ 8!\ 2!\ 8!} = 0.011$$

Cumulative tail probability = 0.011 + 0.003 + 0.007 = 0.021 and is thus significant at an $\alpha = 0.021$ level.

If we perform a similar unpaired analysis of the first period for demonstrating a treatment effect we likewise obtain a significant test at $\alpha = 0.021$ level. Suppose carryover effect would be smaller, e.g., $p_A = 0.8$, $p_B = 0.0$, $p_C = 0.2$. Then the test for treatment effect would yield an even better result:

$$\text{Point probability for carryover effect} = \frac{29!\ 8!\ 10!\ 10!}{20!\ 2!\ 8!\ 10!\ 0!} = 0.004$$

Cumulative tail probability = 0.004 + 0.001 + 0.003 = 0.008.

So, in crossovers with a binary response and a negative result, it does make sense to test for carryover effect by comparing the two periods with the less effective treatment modalities. If a significant test is demonstrated, we obviously will find a significant difference at a similar or even lower level of significance when taking the 1st period for estimating the difference between treatment A and B. Thus, it would seem appropriate for our purpose to disregard the data of the 2nd period in this particular situation (although the 2nd period might still provide interesting information).

Table 4. Example

	Period I		**Period II**	
	Treatment	**Probability of treatment success**	**Treatment**	**Probability of treatment success**
Group I (n = 10)	A	$p_A = 0.8$	B	$p_B = 0.2$
Group II (n = 10)	B	$p_B = 0.2$	A	$p_C = 0.2$

6. DISCUSSION

The power of crossover studies is frequently reduced by carryover effect. This is particularly so when a group that is treated with an effective treatment first, is then treated with an ineffective treatment or placebo second. In studies with a quantitative response this very effect may cause severe underestimation of the treatment effect.[1] Studies with a binary response are, however, different from studies with a quantitative response in that they are mostly designed to answer whether a treatment has any effect rather than what size such effect does have. One might contend, therefore, that underestimation in such studies is not that important as long as the null hypothesis of no treatment effect doesn't have to be erroneously accepted. We demonstrate that in crossovers with a binary response and significant carryover effect the power of testing the treatment effect remains substantial even so. This would imply that routinely testing for carryover effects in such studies is not necessary as long as the result of the treatment comparison is positive. When a study is negative it does make sense, however, to test for carryover effect by comparing p_A versus p_C (table 1).

When p_A is significantly different from p_C, we assume that there is carryover effect in group II. In this situation a parallel-group analysis of period I (p_A versus p_B) can effectively be used for the purpose of demonstrating a treatment effect. It will provide a significant difference at the same or even a lower level of significance than the test for carryover effect. This is so, because when carryover effect is maximal, p_B equals p_C. The difference between p_B and p_A will, therefore, be at least as large as the difference between p_C and p_A but probably larger. Therefore, no further test for treatment effect seems to be required for our purpose and it seems appropriate that the results of the 2nd period be disregarded.

Considering that the problem of carryover effects influence in crossover trials with a binary response may not be too hard to handle, we may as well shift our standard of choosing this particular trial design somewhat, and make use of its additional advantages more frequently. The design is, e.g., particularly powerful for the study of rapid relief of symptoms in chronic disease where the long-term condition of the patient remains fairly stable.[16] This is so, because between-subject variability is not used in a within-subject comparison. Also, we can make use of positive correlations between the treatment modalities tested, because the statistical power of testing treatment comparisons with a positive correlation can be largely enhanced by within-subject comparisons.[17] Furthermore, none of the patients in the trial has to be treated throughout the trial with a less adequate dose or placebo, which is why a crossover raises usually less ethical problems than does a parallel-group study where one group is treated with a placebo or less adequate dosage throughout the trial. Also, we have the advantage that patients can express their own opinions about which of the treatments they personally prefer. This is especially important with subjective variables, such as pain scores.

Furthermore, not so large a group is required because of within-subject comparisons, which facilitates the recruitment procedure and reduces costs. Finally,

double-blinding cannot be effectively executed in self-controlled studies without some kind of crossover design.
In summary:

1. Crossover studies with a binary response and positive results do not have to be tested for carryover effects.
2. If such studies have a negative result, testing for carryover effect does make sense.
3. If a carryover effect is demonstrated, the treatment results should be analyzed in the form of a parallel-group study of the 1st period.

7. CONCLUSIONS

The two-period crossover trial has the evident advantage that by the use of within-patients comparisons, the usually larger between-patient variability is not used as a measuring stick to compare treatments. However, a prerequisite is that the order of the treatments does not substantially influence the outcome of the treatment. Crossover studies with a binary response (such as yes / no or present / absent), although widely used for initial screening of new compounds, have not previously been studied for such order effects. In the present paper we use a mathematical model based on standard statistical tests to study to what extent such order effects, otherwise called carryover effects, may reduce the power of detecting a treatment effect. We come to the conclusion that in spite of large carryover effects the crossover study with a binary response remains a powerful method and that testing for carryover effects makes sense only if the null-hypothesis of no treatment effect cannot be rejected.

8. REFERENCES

1. Cleophas TJM: A simple analysis of carryover studies with one-group interaction. Int J Clin Pharmacol Ther 1995; 32: 322-28.
2. Cleophas TJ: Underestimation of treatment effect in crossover trials. Angiology 1990; 41: 855-64.
3. Brown BW: The crossover experiment for clinical trials.Biometrics 1980;36:69-79
4. Barker M, Hew RJ, Huitson A, Poloniecki J: The two-period crossover trial. Bias 1982; 9: 67-112.
5. Louis TA, Lavori PW, Bailar JC, Polansky M: Crossover and self-controlled design in clinical research. N Engl J Med 1984; 310: 24-31.
6. Willan AR, Pater JL: Carryover and the two-period clinical trial. Biometrics 1986; 42: 593-9.
7. Packer M: Combined beta-adrenergic and calcium entry blockade in angina pectoris. N Engl J Med 1989; 320: 709-18.
8. Fleiss JL: A critique of recent research on the two-treatment crossover design. Control Clin Trials 1989; 10: 237-43.
9. Freeman PR: The performance of the two-stage analysis of two-treatment, two-period crossover trials. Stat Med 1989; 8: 1421-32.

10. Senn S. Crossover trials in clinical research. Wiley & Sons, Chicester, 1993.
11. Hills M. Armitage P: The two-period crossover trial. Br J Clin Pharmacol 1979; 8: 7-20.
12. Gart JJ: An exact test for comparing matched proportions in crossover designs. Biometrika 1969; 56: 57-80.
13. Prescott RJ: The comparison of success rates in crossover trials in the presence of an order effect. Appl Stat 1981; 30: 9-15.
14. Fidler V: Change-over clinical trials with binary data: mixed model-based comparisons of tests. Biometrics 1984; 40: 1063-79.
15. Bavry JH: Design Power (TM). Scientific software Inc., 1988, Hillsdale, New Jersey.
16. Cleophas TJM, Tavenier P. Clinical trials of chronic diseases. J Clin Pharmacol 1995; 35: 594-8.
17. Cleophas TJM, Tavenier P. Fundamental issues of choosing the right type of trial. Am J Ther 1994; 1: 327-32.

CHAPTER 25

CROSS-OVER TRIALS SHOULD NOT BE USED TO TEST TREATMENTS WITH DIFFERENT CHEMICAL CLASS

1. INTRODUCTION

So many unpredictable variables often play a role in clinical trials of new medical treatments that a trial without controls has become almost unconceivable. Usually, a parallel-group design is used: with every patient given a new therapy, a control patient is given standard therapy or a placebo. For the study of reversible treatments of chronic stable conditions with responses that can be measured on relatively short notice a cross-over design can be chosen: a single patient receives both new therapy and a standard therapy or placebo. Of course, we have to be fairly sure that carryover effects of one treatment period carrying on into the other or time effects are negligible. But then the cross-over design has the advantage that it eliminates between-subject variability of symptoms in a treatment comparison. And this makes the design sensitive, particularly with conditions where between-subject variability is notoriously large, e.g., angina pectoris and many other pain syndromes.

In 1965 the biostatistician James Grizzle[1] gave uniform guidelines for the cross-over design, and it was he who first recognized the problem of negative correlations between treatment responses that may endanger the validity of the cross-over design. In his example two completely different treatments (A = ferrous sulphate and B = folic acid) were tested for their abilities to increase hemoglobin (Figure 1). Obviously, there was an inverse correlation between the two treatments: ferrous sulphate was only beneficial when folic acid was not, and so was folic acid when ferrous sulphate was not. Although the mean result of ferrous sulphate treatment was 1.7 mmol different from that of folic acid which is quite a difference, it did not reach statistical significance ($p = 0.12$). This was probably due to the significant negative correlation in the treatment comparison. How a negative correlation reduces the sensitivity of a paired comparison can be explained as follows:

t = mean result / pooled SEM.
where pooled SEM = pooled standard error of the mean

the formula for pooled SEM is:
$(\text{pooled SEM})^2 = \text{SEM}_1^2 + \text{SEM}_2^2 - 2\, r\, \text{SEM}_1.\text{SEM}_2$

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 309–318.

where SEM_1 and SEM_2 are standard errors of the mean of separate treatments and r = correlation coefficient.

When we assume $SEM_1 = SEM_2 = SEM$, then
$(\text{pooled SEM})^2 = (1\text{-r})\ 2\ SEM^2$

If r would have been 0 instead of - 0.49 (figure 1) the t - value of this comparison would have been $\sqrt{(1-r)} = \sqrt{1.49} = 1.225$ larger than the present t - value and the treatment comparison would have reached statistical significance.

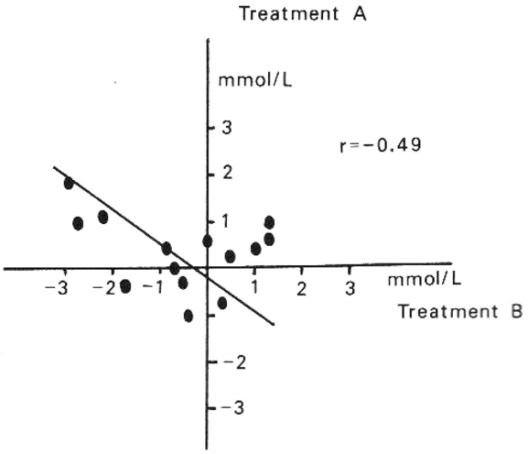


Figure 1. Two completely different treatments (A=ferrous sulphate and B=folic acid) were tested for their abilities to increase hemoglobin. There was an inverse correlation between the two treatments: ferrous sulphate was only beneficial when folic acid was not, and so was folic acid when ferrous sulphate was not. Although the mean result of ferrous sulphate treatment was 1.7 mmol different from that of folic acid, the difference did not reach statistical significance (p = 0.12). This was probably due to the negative correlation in the treatment comparison (Grizzle 1965).[1]

We currently are aware that ferrous sulphate and folic acid are treatments with a totally different chemical class / mode of action. And so, although both of the compounds improve hemoglobin, certainly nobody nowadays would use the compounds in a treatment comparison anymore. However, we continue to compare many other treatments from different classes of drugs all the time, even if we know that their mode of action is totally different, e.g., beta-blockers are compared with calcium channel blockers or nitrates for the treatment of angina pectoris. Compounds from different chemical classes are compared for the treatment of

hypertension, Raynaud's phenomenon, cardiac arrhythmias, chronic obstructive pulmonary disease and many more conditions.

The current chapter shows that it is not correct to use a cross-over design for testing such kind of treatment comparisons because of the risk of negative correlations between treatment responses, and thus of a flawed study. We will test this hypothesis in a non-mathematical way by giving examples in which a cross-over design should NOT have been used. Also, we will estimate the size of the problem by reviewing hypertension trials published for their design in relation to the type of treatment comparison. A more mathematical approach of the problems of negative correlations can be found elsewhere.[2]

2. EXAMPLES FROM THE LITERATURE IN WHICH CROSS-OVER TRIALS ARE CORRECTLY USED

Cross-over trials generally have a strong positive correlation between treatment responses for 2 reasons. First, this is so, because one subject is used to the comparison of two treatments. Second, in controlled clinical trials the new treatment may be a slight modification of the standard or be equivalent to it with the addition of a new component. In this situation there is a positive correlation between the response to the new treatment and the standard treatment: treatment 1 performs highly when treatment 2 does so.

Table 1 gives 7 examples of cross-over studies where compounds from the same chemical class / mode of action are compared. E.g., two beta-adrenergic agonists, two calcium channel blockers, two beta-blockers, two different dosages of the same compound are compared. Such comparisons should have a strong positive correlation, and the table shows that this is so. Correlation coefficients calculated from the data were consistently positive. These studies were appropriately performed in the form of a cross-over study. The cross-over design provided extra sensitivity by accounting for the positive correlation. A parallel-group study would have lacked the extra sensitivity.

Table 1. Examples from the literature in which cross-over trials are correctly used

	Treatment	Efficacy[0] (mean ± SEM)	P-value	Correlation Coefficient[#]
1. Angiology 1985; 36: 219-26 n = 12	beta-adrenergic agonist alpha-adrenergic antagonist with beta-agonistic property	22.7 ± 0.5 27.7 ± 1.0	<0.01	r = + 0.56
2. Lancet 1986; ii: 189-92 n = 6	Platelet activating factor its precursor	-1.5 ± 1.0 +0.2 ± 1.0	< 0.001	r = + 0.66
3. Lancet 1986; ii: 740-1 n = 7	cholesterol lowering drug A cholesterol lowering drug B	42 ± 12 50 ± 12	< 0.05	r = + 0.20
4. Lancet 1987; i: 647-52 n = 40	high alcohol intake low alcohol intake	143 ± 5 137 ± 5	< 0.01	r = + 0.41
5. Lancet 1987; ii: 650-3 n = 20	atenolol labetalol	74.3 ± 4.5* 79.9 ± 7.2	< 0.01	r = + 0.39
6. Br Heart J 1993; 70: 252-8 n = 18	gallopamil nifedipine	29.9 ± 11.0 49.7 ± 26.8	< 0.0001	r = + 0.56
7. Int J Clin Pharmacol Ther 1997; 35: 514-8 n= 8	amlodipine felodipine	1.58 ± 0.32 4.43 ± 1.86	<0.001	r = + 0.65

[0] Denotes in study 1 finger temperature after finger cooling (^oC), study 2 bronchial responsiveness to methacholine (doubling dilutions), in study 3 plasma level of HDL-cholesterol (mg/dl), in study 4 systolic blood pressure (mm Hg), in study 5 heart rate (beats / min), in study 6 QRS voltage (% of standardized maximum), in study 7 peak-trough ratio.

[#] Correlation coefficient (r) was calculated using t - statistic: p - values were turned into t – values after adjustment for the degrees of freedom, and r was calculated using the formula for the pooled standard error of the mean (SEM): (pooled SEM)^2 = SEM_1^2 + SEM_2^2 - 2 r $SEM_1.SEM_2$

***** For the paired analysis two-sided ANOVA was used which for two groups of paired data yields the same results as a paired t - test, however.

3. EXAMPLES FROM THE LITERATURE IN WHICH CROSS-OVER TRIALS SHOULD NOT HAVE BEEN USED

In trials with completely different treatments patients tend to fall apart into different populations: those who respond better to treatment 1 and those who do so to treatment 2. For example, patients with angina pectoris irresponsive to beta-blockers may respond either to calcium channel blockers or nitrates. Also, hypertension, Raynaud's phenomenon, different types of cardiac arrhythmias and chronic obstructive pulmonary disease are known to be conditions where a non-response to a particular compound is frequently associated with an excellent response to a completely different compound. These are examples of situations in which a strong negative correlation may exist. This may be even so with self-controlled studies that otherwise are more likely to have a positive correlation because one subject is used to the comparison of two treatments. As demonstrated above the problem with negative correlations in a cross-over study is lack of sensitivity: the pooled SEM is approximately $\sqrt{(1-r)}$ times larger with a negative correlation than it would have been with a zero correlation (parallel-group study), and this reduces the probability level of testing, and, thus, produces erroneously negative studies. The examples in Table 2 show that the problem can be readily detected in the literature. All of these studies were negative, and this was presumably so because of the negative correlation coefficient between treatment responses. Had they been performed in the form of a parallel-group study, most of them probably would have had a statistically significant effect. At least, when we tested the studies as though they were unpaired, in most of them p - values of 0.05 or less were obtained.

Table 2. Examples from the literature in which cross-over trials should NOT have been used

	Treatment	Efficacy[0] (mean ± SEM)	P-value	Correlation coefficient[#]
1. Lancet 1986; i: 997-1001 n = 20	NSAID with renal NSAID without renal prostaglandin synthesis	127 ± 3 131 ± 3	n.s.	r = - 0.29
2. N Engl J Med 1986; 314: 1280-6 n = 12	tolazolin insulin	140 ± 34 112 ± 15	n.s.	r = - 0.30
3. N Engl J Med 1986; 315: 735-9 n = 11	beta-adrenergic agonist anticholinergic agent	42 ± 18 25 ± 14	n.s.	r = - 0.30
4. Br J Clin Pharmacol 1991; 31: 305-12 n = 38	xamoterol enalapril	80.1 ± 2.6 75.1 ± 1.6	n.s.	r = - 0.25
5. Br J Clin Pharmacol 1991; 32: 758-760 n = 6	nitroprusside bradykinine	13 ± 5 91 ± 2	n.s.	r = - 0.42
6. Curr Ther Res 1991; 49: 340-50 n = 42	nifedipine captopril	14.0 ± 3.6 6.7 ± 2.1	n.s.	r = - 0.46
7. Eur J Gastroenterol Hepat 1993; 5: 627-9 n = 18	atenolol nifedipine	3.9 ± 0.2 2.9 ± 0.3	n.s.	r = - 0.70

SEM = standard error of the mean, n.s. = not significant.

[0] Denotes in study 1 systolic blood pressure (mm Hg), in study 2 plasma glucose level (mg/dl), in study 3 forced expiratory volume in one second (% change from baseline), in study 4 diastolic blood pressure (mm Hg), in study 5 plasma ureum (mmol/l), in study 6 fall in mean blood pressure (mm Hg), in study 7 oesophageal sphincter pressure (mm Hg).

[#] Correlation coefficient (r) was calculated using t - statistic: p - values were turned into t - values for the degrees of freedom, and r was calculated using the formula for the pooled standard error of the mean (SEM): $(\text{pooled SEM})^2 = SEM_1^2 + SEM_2^2 - 2\, r\, SEM_1.SEM_2$.

4. ESTIMATE OF THE SIZE OF THE PROBLEM BY REVIEW OF HYPERTENSION TRIALS PUBLISHED

The above examples indicate the existence of a potential problem with negative correlations in cross-over trials. However, they do not answer how prevalent the problem is. In order to address this question we assessed the double blind randomized hypertension trials listed in Cardiology Dialogue 1994.[3] Hypertension treatments frequently have pharmacologically completely different modes of action: diuretics reduce blood pressure by volume depletion, beta-blockers and calcium channel blockers / angiotensin converting enzyme inhibitors do so by reducing cardiac output and peripheral resistance respectively. Of 73 randomized controlled trials (Table 3) a significantly smaller percentage of cross-over than of parallel-group studies compared treatments with a totally different chemical class / mode of action (for example, diuretic versus vasodilator, or beta-blocker versus vasodilator etc, 27 versus 72%, $P<0.001$). Apparently, the scientific community has some intuition of doing the right thing at the right time: in 73% of the cases the cross-over design was correctly used. Nonetheless, in 4 (27%) of the cases this was not so. Two of these studies were not able to reject the null-hypothesis of no effect and the other two would probably have been more sensitive, had they been performed in the form of a parallel-group study.

Table 3. Double blind randomized hypertension trials listed in the 1994 volume of Cardiology Dialogue[4]

	parallel-group studies		cross-over studies	
	N	different treatments (%)	N	different treatments (%)
Am J Cardiol	3	2	2	1
J Am Coll Cardiol	1	0		
Am J Hypertens	7	5	1	0
Curr Ther Res	5	2	1	0
Clin Med			1	0
NEJM	2	0		
Clin Exp Hypertens	1	0	2	0
J Human Hypertens	7	6	2	0
Br J Clin Pharmacol	3	3	1	1
Cardiovasc Drug Ther			1	1
Clin Lab Invest	1	1		
Herz Kreislauf	2	1		
Zeitschr Kardiol	1	1		
J Cardiovasc Pharmacol	4	3		
J Clin Pharmacol	3	2	1	0
Clin Ther			1	0
Clin Pharmacol Ther	2	2		
Cardiol	4	3		
J Int Med	1	1		
Eur J Clin Pharmacol	1	1		
Hypertens	1	1		
Arch Int Med	1	1		
B J Clin Pract			1	1
Clin Pharmacol Res	1	1		
JAMA	1	1		
Postgrad Med	1	1		
Drug Invest			1	0
Total numbers	53	38 (72%)	15	4 (27%)

5. DISCUSSION

The current chapter shows that clinical trials comparing treatments with a totally different chemical class / mode of action are at risk of negative correlation between treatment responses. Such negative correlations have to be added to the standard errors in a cross-over trial, thus reducing the sensitivity of testing differences, making the design a flawed method for evaluating new treatments. The examples suggest that the phenomenon of negative correlations is not uncommon in practice, and that it should be taken into account when planning drug research.

The mechanism of between-group disparities in drug response is currently being recognized in clinical pharmacology, and is, in fact, the main reason that in treatment protocols the principle of stepped care is being replaced by individualized

care.[4] However, when it comes to research, clinicians and clinical pharmacologists are still unfamiliar with the problems this issue raises and virtually never take account of it. The recognition of between-group disparities in drug response also implies that negative correlations in a treatment comparison are routinely tested, and that a cross-over design is not always appropriate.

So far, statisticians have assumed that a negative correlation in cross-over studies was virtually non-existent, because one subject is used for comparison of two treatments. For example, Grieve recently stated one should not contemplate a cross-over design if there is any likelihood of correlation not being positive.[5] The examples in the current paper show, however, that with completely different treatments, the risk of a negative correlation is a real possibility, and that it does give rise to erroneously negative studies. It makes sense, therefore, to restate Grieve's statement as follows: one should not contemplate a cross-over design if treatments with a totally different chemical class / mode of action are to be compared.

At the same time, however, we should admit that the cross-over design is very sensitive for comparing treatments of one class and presumably one mode of action. The positive correlation in such treatment comparisons adds sensitivity, similarly to the way it reduces sensitivity with negative correlations: the pooled SEM is approximately $\sqrt{(1-r)}$ times smaller with positive correlation than it would have been with a zero correlation (parallel-group study), and this increases the probability level of testing accordingly. This means that the cross-over is a very sensitive method for evaluating studies with presumable positive correlation between treatment responses, and that there is, thus, room left for this study design in drug research.

6. CONCLUSIONS

Comparisons of treatments with totally different chemical class / mode of action are at risk of a negative correlation between treatment responses: patients tend to fall apart into different populations, those who respond better to treatment 1 and those who do so to treatment 2. The cross-over design is flawed when this phenomenon takes place. The objective of this chapter was to assess whether this flaw is prevalent in the literature.

Fourteen randomized controlled cross-over studies were assessed for correlation levels in relation to their type of treatment comparison. Correlation coefficient (r) was calculated using T-statistic: P-values were turned into T-values for the degrees of freedom, and r was calculated using the formula for the pooled standard error of the mean (SEM): $(\text{pooled SEM})^2 = SEM_1^2 + SEM_2^2 - 2\, r\, SEM_1.SEM_2$. Randomized controlled hypertension trials of 1994 were listed for study design in relation to type of treatment comparison.

Cross-over studies comparing treatments with a totally different chemical class / mode of action were frequently negative, and this was, obviously, due to their negative correlation between treatment responses. Cross-over studies comparing similar treatments had frequently a positive correlation, and this added extra sensitivity to the treatment comparison. Twenty-seven percent of the cross-over hypertension studies compared completely different treatments, and these studies should, therefore, not have been performed in the form of a cross-over study.

Cross-over trials lack sensitivity to test one treatment against another treatment with a totally different chemical class / mode of action, and should, therefore, not be used for that purpose. In contrast, they are, particularly, sensitive to compare treatments from one chemical class / with one mode of action. It is hoped that this chapter affects the design of future crossover trials.

7. REFERENCES

1. Grizzle JE: The two-period change-over design and its use in clinical trials. Biometrics 1965; 22: 467-80.
2. Cleophas TJ: Between-group disparities in drug response. In: Human Experimentation Boston, Kluwer Academic Publishers, 1999; 48-56.
3. Cardiology Dialogue: Edited by Rapid Literature Service. Cologne, Germany: Limbach GMBH, 1994.
4. Nies AS, Spielberg SP: Individualization of drug therapy. In: Hardman JL, Limbird LE, eds. Goodman and Gilman's Pharmacological Basis of Therapeutics New York: McGraw-Hill, 1996: 43-63.
5. Grieve AP: Bayesian analysis of two-treatment crossover studies. Stat Meth Med Res 1994; 3: 407-29.

CHAPTER 26

QUALITY-OF-LIFE ASSESSMENTS IN CLINICAL TRIALS

1. INTRODUCTION

Less than 10 years ago the scientific community believed that quality of life (QOL) was part of the art of medicine rather than the science of medicine. In the past few years index methods have been developed and have proven to be sensitive and specific to assess patients' health status not only on a physical, but also on a psychological and social base. We increasingly witness that QOL is implemented in the scientific evaluation of medicine. However, major problems with QOL assessments so far, include the contributing factor patients' opinion, which is very subjective and, therefore, scientifically difficult to handle, and, second, the low sensitivity of QOL-questionnaires to reflect true changes in QOL. The Dutch Mononitrate Quality Of Life (DUMQOL) Study Group has recently addressed both problems. In their hands, the patients' opinion was a consistent and statistically independent determinant of QOL in patients with angina pectoris. The problem of low sensitivity of QOL-assessments could be improved by replacing the absolute score-scales with relative ones, using for that purpose odds ratios of scores. The current chapter reviews the main results of this so far only partly published research[1,2] from the Netherlands.

2. SOME TERMINOLOGY

QOL battery	A questionnaire large enough to adequately address important domains of QOL.
Domains of QOL	Physical, psychological, and social areas of health seen as distinct and important to a person's perception of QOL.
Items	Items, otherwise called questions, constitute a domain, e.g., the DUMQOL-questionnaire for angina pectoris, consists of respectively 8, 7, and 4 questions to assess the domains (1) mobility, (2) somatic symptoms, and (3) psychological distress.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 319–329.

Absolute score scales	For every item the individual response is scored on a (linear) scale. Mean of scores a group of patients are calculated. Mean domain scores are calculated as overall means of the latter mean scores.
Relative score scales	The same procedure. However, results are reported in the form of odds ratios.
Odds ratios	Mean of the domain scores in patients with a particular characteristic / mean of the domain scores in patients without this particular characteristic.
Validated QOL batteries	This is controversial. QOL batteries are diagnostic tests, and validation of any diagnostic test is hard to accomplish without a gold standard for comparison. Surrogate validation is sometimes used: actual QOL scores are compared with scores expected based on levels of morbidity.
Internal consistency of domain items	There should be a strong correlation between the answers given to questions within one domain: all of questions should approximately predict one and the same thing. The level of correlation is expressed as Cronbach's alpha: 0 means poor, 1 perfect relationship.
Cronbach's alpha	$\text{alpha} = \frac{k}{(k-1)} \cdot (1 - \sum \frac{s_i^2}{s_T^2})$ k= number of items s^2_i = variance of ith item s^2_T = variance of total score obtained by summing up all of the items
Multicollinearity	There should not be a too strong correlation between different domain scores because different domains predict different areas of QOL. A Pearson's correlation coefficient > 0.90 means the presence of multicollinearity and, thus, of a flawed multiple regression analysis.
Pearson's correlation coefficient (r)	$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$

Sensitivity of QOL assessment	Sensitivity or precision means ability of the measurement to reflect true changes in QOL.
QOL estimator	Mean (or pooled) result of the data from a single domain.
Index methods	Index methods combine the results of various domains of a QOL battery to provide an index for overall QOL.

3. DEFINING QOL IN A SUBJECTIVE OR OBJECTIVE WAY?

In 1992 Brazier et al[3] validated the Short Form (SF)-36 health survey questionnaire of Stewart[4], a self-administered questionnaire, addressing any aspects that, according to the designer, might be important to the patients' QOL. However, at each item in the questionnaire, the question "is it important to you?" was missing. In 1994 Gill and Feinstein[5] in their "Critical appraisal of quality of life assessments" emphasized that, from their personal experience in patient care, they believed that QOL, rather than a description of health status, should describe the way patients perceive their health status. One year later Marquis et al[6] designed a questionnaire for patients with angina pectoris based on psychological factors, in addition to clinical symptoms, and concluded that the former is probably a better predictor of QOL than the latter. In subsequent years QOL assessments increasingly allowed for patients giving their own opinion, in addition to patients answering questions about health status. However, the latter was consistently given more weight than the former. For example, Testa and Simonson[7] allowed for one such question out of 6 questions in each QOL-domain giving the question just about 1/6 of the total weight in various domains. The problem with the subjective approach to QOL, as recently pointed out by Thompson et al[8], is that it is difficult to match with the accepted rule that scientific data should be objective. In addition, the patients' opinion may be a variable so unpredictable, that it cannot be applied as a reliable measure for clinical assessment of groups of patients. So far, the concept that the patients'opinion is a relevant variable in the assessment of QOL has never been proven to be true. In order to test this issue the DUMQOL Study Group has recently completed some relevant research.

4. THE PATIENTS' OPINION IS AN IMPORTANT INDEPENDENT-CONTRIBUTOR TO QOL

The DUMQOL Study Group used the validated form of Stewart's SF-36 Questionnaire for the purpose of scoring QOL[3], and the DUMQOL-50 questionnaire for scoring psychological distress and health status according to the patients' judgment.[9] The patients' opinion (patients were requested to estimate the overall amount of his/her QOL as compared to patients they knew with a similar condition) and health status according to the physicians' judgement (the physician was requested to estimate the patients' health status) were scored like the others on 5 point-scales. Internal consistency and retreatment reliability of the test-battery was adequate with Cronbach's alpha 0.66. Table 1 shows the results from a cohort of 82 outpatient-clinic

Table 1. Correlation matrix to assess multicollinearity in the data, Pearson's correlation coefficient are given (r)

	Patients' opinion	psychological distress	health status patients' judgment	health status physicians' judgment
Psychological distress	0.35			
Health status Patients' judgment	0.36	0.30		
Health status physicians' judgment	0.42	0.41	0.48	
Quality of life	0.42	0.58	0.43	0.27

R < 0.20 weak correlation; 0.20 < r < 0.40 moderate correlation; r > 0.40 strong correlation.

Table 2. Stepwise multiple regression analysis of the associations of various (dependent) predictors on QOL in patients with angina pectoris

	beta	t	p - value
Psychological distress	0.43	4.22	0.000
Patients' opinion	0.22	2.19	0.032
Health status (patients' judgment)	0.19	1.88	0.071
Health status (physicians' judgment)	0.11	0.16	0.872

beta= standardized partial correlation coefficient.

patients with stable angina pectoris. Obviously, QOL was strongly associated with the patients' opinion. In none of the comparisons were adjustment for multicollinearity required (Pearson's correlation coefficient >0.9). Table 2 shows that psychological distress was the most important contributor to QOL. Also, the patients' opinion significantly contributed to QOL. Physical health status according to the patients' judgment only made a borderline contribution, while the physicians' judgment was not associated with QOL at all. These data strongly support the relevance of the patients' opinion as an important independent-contributor to QOL.

5. LACK OF SENSITIVITY OF QOL-ASSESSMENTS

Sensitivity defined as ability of the measurement to reflect true changes in QOL is frequently poor in QOL assessments.[10] A well-established problem with QOL scales is their inconsistent relationship between ranges of response and true changes in QOL.[7] A good example of this problem is the physical scale of the SF-36 questionnaire. It ranges from 0 to 100 points. However, while healthy youngsters may score as high as 95 and topsporters even 100, 60 year-old subjects usually score no better than 20. A patient with angina pectoris may score 5 points. If he would score 10, instead of 5, after the allowance for sublingual nitrates ad libitum, this improvement would equal 5% on the absolute scale of 100 points, which does not seem to be very much. However, on a relative scale this score of 10 points is 100% better than a score of 5 points, and, in terms of improvement of QOL, this difference on the SF-36-scale between 5 and 10 points does mean a world of

difference. It, for example, means the difference between a largely dependent and independent way of life. In this example the low score on the absolute-scale masks important and meaningful changes in QOL. The DUMQOL Study Group took issue with this well-recognized but unsolved phenomenon and performed an odds ratio analysis of patient characteristics in a cohort of 1350 patients with stable angina pectoris. They showed that this approach provided increased precision to estimate effects on QOL estimators.

6. ODDS RATIO ANALYSIS OF EFFECTS OF PATIENT CHARACTERISTICS ON QOL DATA PROVIDES INCREASED PRECISION

Table 3 gives an overview of effects of patient characteristics on QOL estimators in 1350 patients with stable angina pectoris. Results are presented as odds ratios. The odds ratio presents the relative risk of QOL difficulties and is defined as the ratio between mean domain score of patients with a particular characteristic and that of patients without this particular characteristic.

Table 3. Stable angina pectoris: effects of patient characteristics on quality of life estimators. Odds ratios and 95% confidence intervals are given

	Mobility difficulties	Pain in general	Early morning pain	Psychological distress	Chest pain	Patient satisfaction
Gender (females/males)	2.5 (1.8-3.3)[c]	2.0 (1.3-3.0)[c]	1.7 (0.6-4.7)	1.3 (0.9-2.0)	2.1 (1.1-3.9)[b]	0.8 (0.3-1.9)
Age (>68/<86 years)	1.4 (1.2-1.5)[b]	1.0(0.9-1.1)	0.9 (0.9-1.0)	1.0 (0.9-1.0)	1.0 (0.9-1.0)	1.0 (0.9-1.0)
NYHA (III-and-IV / II-and-I)	5.6 (4.8-6.6)[c]	2.8 (2.1-3.5)[c]	46.8 (26.3-83.1)[c]	4.4 (3.5-5.5)[c]	37.2 (23.4-58.9)[c]	0.6 (0.4-1.1)
Smoking yes/no	0.8 (0.5-1.1)	1.3 (0.8-2.1)	12.9 (3.0-56.2)[c]	3.2 (2.0-5.2)[a]	0.5 (0.2-1.2)	5.8(2.1-15.8)[b]
Cholesterol yes/no	0.9 (0.7-1.3)	1.4 (0.3-2.0)	1.3 (0.5-3.4)	1.8 (1.2-2.8)[a]	1.8 (0.9-3.4)	1.1 (0.5-2.6)
Hypertension yes/no	0.3 (0.2-0.4)[a]	0.5 (0.3-0.7)[a]	0.7 (0.2-0.9)[a]	0.3 (0.2-0.4)[b]	0.5 (0.3-0.9)[a]	1.7 (0.7-4.1)
Diabetes yes/no	2.2 (1.5-3.1)[a]	1.1 (0.6-1.9)	9.1 (3.0-28.2)[c]	2.0 (1.1-3.7)[a]	1.8 (0.7-4.6)	1.1 (0.3-4.2)
Arrhythmias yes/no	2.9 (2.0-4.1)[b]	1.3 (0.7-2.1)	3.6 (1.3-10)[a]	3.2 (1.9-5.4)[a]	10.2 (4.5-23.4)[b]	1.2 (0.4-3.7)
PVD yes/no	11.0 (7.9-15.1)[c]	2.2 (1.4-3.6)[a]	1.1 (0.7-1.7)	2.6 (1.5-4.5)[a]	1.0 (0.4-2.2)	8.3 (2.7-25.7)[b]
Beta-blockers yes/no	0.8 (0.7-0.9)[a]	0.8 (0.5-1.1)	1.7 (0.7-4.0)	0.9 (0.6-1.2)	1.3 (0.7-2.2)	3.2 (1.5-6.9)[b]
Calcium channel blockers yes/no	1.5 (1.2-1.9)[a]	1.3 (0.9-1.8)	3.2 (1.5-6.6)[a]	2.0 (1.4-2.9)[a]	6.0 (3.4-10.7)[b]	6.5 (3.0-13.8)[a]
Sublingual nitrates yes/no	2.6 (2.1-3.3)[c]	3.0 (2.2-4.2)[c]	1.0 (0.7-1.4)	3.1 (2.5-4.3)[c]	7.1 (4.2-12.0)[c]	3.4 (1.6-6.9)[c]

Quality of life domains were estimated using a questionnaire based on the Medical Outcomes Short-Form 36 Health Survey and the Angina Pectoris Quality of Life Questionnaire. Results are given as odds ratios = mean domain scores in patients with characteristic /mean domain scores in patients without characteristic. PVD = peripheral vascular disease; NYHA= New York Heart Association Angina Class; a = $P < 0.05$; b = $P < 0.01$; c = $P < 0.001$.

The procedure readily identifies categories of patients that, obviously, have poor QOL scores. E.g.,

1. Increased QOL-difficulties were observed in patients with advanced New York Heart Association (NYHA) anginal class: the higher the anginal class the larger the risk of mobility difficulties, pain, chest pain, anginal pain, and distress.
2. The risk of mobility difficulties was increased in patients with diabetes mellitus, arrhythmias, and peripheral vascular diseases.
3. Patients using sublingual nitrates (and thus presumably very symptomatic) reported more (severe) mobility difficulties, pain, chest pain, and psychological distress.
4. Female patients reported more (severe) mobility difficulties, pain, anginal pain, and distress than their male counterparts.
5. The risk of mobility difficulties increased with age, but, in contrast, elderly patients reported less pain, anginal pain, and distress.

The above categories of patients are, obviously, very symptomatic and should, therefore, particularly benefit from treatments. The beneficial effects of treatments in patients with particular characteristics can be predicted according to the following procedure:

(1) Odds Ratio $_{\text{active treatment / placebo}}$ = mean domain score in patients on active treatment/ mean domain score in patients on placebo.

(2) Odds Ratio $_{\text{characteristic / no characteristic}}$ = mean domain score in patients with particular characteristic / mean domain score in patients without this particular characteristic.

The relative risk of scoring in patients with a particular characteristic if they used active treatment

can be estimated and calculated according to:

(3) Odds Ratio $_{\text{characteristic / no characteristic}}$ x Odds Ratio $_{\text{active treatment / placebo}}$.

Along this line the odds ratio approach to QOL-assessments can be helpful to estimate the effects of cardiovascular drugs on quality of life in different categories of patients with increased precision.

7. DISCUSSION

The medical community is, obviously, attracted to the concept that QOL assessments should pay particular attention to the individual, but, at the same time, it believes in the usefulness of a scientific method to measure QOL.[11] Usually, the objective of a study is not to find the greatest good for a single person but the greatest good for the entire population, moving from an individual perspective to a societal one. Even for quality-of-life measurements, only large clinical studies designed and conducted with rigorous statistical standards allow for a hypothesis to be tested and to offer useful results. Using the patients' opinion as measurement-instrument raises a major problem within this context. The general concept of medical measurements is that measurement-instruments remain constant irrespective of who is using them: a thermometer remains the same whoever's mouth it is placed in. With the patients' opinion this is not so. Rather than true ability, perceived functional ability and willingness to complain is assessed. An assessment tool to reflect the viewpoint of patients is, obviously, a major challenge. Although the medical community expresses sympathy with the latter concept, it expresses doubt about scientific value and even questions whether the patients' opinion is part of medicine at all.[7,8,11] The recent research from the DUMQOL Group shows that the patients' opinion in a standardized way, produces data that are sufficiently homogeneous to enable a sensitive statistical analysis. These data strongly support the relevance of the patients' opinion as an independent contributing factor to QOL. This variable should, therefore, be adequately implemented in future QOL assessments.

A second problem with current QOL-batteries is the inconsistent relationship between ranges of response and true changes in QOL-assessments. This is mainly due to very low (and very high) scores on the absolute-scale, masking important and meaningful changes in QOL. The DUMQOL Study Group showed that this problem can be adequately met by the use of relative rather than absolute scores, and it used for that purpose an odds ratio-approach of QOL scores. This approach provided increased precision to estimate effects on QOL estimators. An additional advantage of the latter approach is that odds ratios are well understood and much in use in the medical community, and that (those) results from QOL research can, therefore, be more easily communicated through odds ratios than through the comparison of absolute scores. For example, "the odds ratio of (severe) mobility difficulties for mononitrate therapy in patients with stable angina is 0.83 ($p < 0.001$)" is better understood than "the mean mobility difficulties score decreased from 1.10 to 1.06 on a scale from 0 to 4 ($p = 0.007$)".

We conclude that recent QOL-research from the DUMQOL Study Group allows for some relevant conclusions, pertinent to both clinical practice and clinical research. QOL should be assessed in a subjective rather than objective way, because the patients' opinion is an important independent contributor to QOL. The comparison of absolute QOL-scores lacks sensitivity to truly estimate QOL. For that purpose the odds ratio approach of QOL scores provides increased precision to estimate QOL.

8. CONCLUSIONS

Two major issues in quality of life (QOL) research include the patients' opinion as a contributing factor in QOL-assessments, and the lack of sensitivity of QOL-assessments. The objective of this chapter was to review results from recent research by the Dutch Mononitrate Quality Of Life (DUMQOL) Study Group relevant to these issues.

Using a test-battery including Stewart's Short Form (SF)-36 Questionnaire and the DUMQOL-50 questionnaire, the DUMQOL Study Group tested the hypothesis that the patients' opinion might be an independent determinant of QOL and performed for that purpose a stepwise multiple regression analysis of data from 82 outpatient clinic patients with stable angina pectoris. Psychological distress was the most important contributor to QOL (beta 0.43, $P<0.0001$). Also, the patients' opinion significantly contributed to QOL (beta 0.22, $p = 0.032$). Physical health status according to the patients' judgment only made a borderline contribution (beta 0.19, $P = 0.71$), while the physicians' judgment was not associated with QOL at all (beta 0.11, $P = 0.87$). Using an Odds ratio approach of QOL scores in 1350 outpatient clinic patients with stable angina pectoris the DUMQOL Study Group assessed the question that relative scores might provide increased precision to estimate the effects of patient characteristics on QOL data. Increased QOL difficulties were observed in New York Heart Association Angina Class (NYHA) III-IV patients, in patients with comorbidity, as well as in females and elderly patients. Odds ratios can be used in these categories to predict the benefit from treatments. We conclude that recent QOL-research of the DUMQOL Study Group allows for conclusions relevant to clinical practice. QOL should be defined in a subjective rather than objective way. The patients' opinion is an important independent contributor to QOL. The comparison of absolute QOL-scores lacks sensitivity to truly estimate QOL. The odds ratio approach of QOL scores provides increased precision to estimate QOL.

9. REFERENCES

1. Frieswijk N, Buunk BP, Janssen RM, Niemeyer MG, Cleophas TJ, Zwinderman AH. Social comparison and quality of life: evaluation in patients with angina pectoris. Cardiogram 2000; 16: 26-31.
2. Zwinderman AH, Niemeyer MG, Kleinjans HA, Cleophas TJ. Application of item response modeling for quality of life assessments in patients with stable angina pectoris. In: Clinical Pharmacology, EDS Kuhlman J, Mrozikiewicz A, Zuckschwerd Verlag, New York, 1999, pp 48-56.
3. Stewart AL, Hays RD, Ware JE. The MOS short form general health survey. Med Care 1988; 26: 724-35.
4. Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, Westlake L, Validating the SF-36 health survey questionnaire: new outcome measure for primary care. Br Med J 1992; 305: 160-4.

5. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. J Am Med Assoc 1994; 272: 619-26.
6. Marquis P, Fagol C, Joire JE. Quality of life assessment in patients with angina pectoris. Eur Heart J 1995; 16: 1554-59.
7. Testa MA, Simonson DC. Assessment of quality-of-life outcomes. N Engl J Med 1996; 334: 835-40.
8. Thompson DR, Meadows KA, Lewin RJ. Measuring quality of life in patients with coronary heart disease. Eur Heart J 1998; 19: 693-5.
9. Niemeyer MG, Kleinjans HA, De Ree R, Zwinderman AH, Cleophas TJ, Van der Wall EE. Comparison of multiple dose and once-daily nitrate therapy in 1350 patients with stable angina pectoris. Angiology 1997; 48: 855-63.
10. Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 Health survey: manual and interpretation guide. Boston. The Health Institute, New England Medical Center, 1993.
11. Albert SM, Frank L, Mrri R, Hylandt, Apolone G, Leplége A. Defining and measuring quality of life in medicine. J Am Med Assoc 1998; 279: 429-31.

CHAPTER 27

STATISTICAL ANALYSIS OF GENETIC DATA

1. INTRODUCTION

In 1860, the benchmark experiments of the monk Gregor Mendel led him to propose the existence of genes. The results of Mendel's pea data were astoundingly close to those predicted by his theory. When we recently looked into Mendel's pea data and performed a chi-square test, we had to conclude the the chi-square value was too small not to reject the null-hypothesis. this would mean that Mendel's reported data were so close to what he expected that we could only conclude that he had somewhat fudged the data (Table 1).

Table 1. Chi-square-distribution not only has a right but also a left tail. We reject the null-hypothesis of no difference with 1 degree of freedom if chi-square is larger than 3.84 or smaller than 0.004. In Mendel's data frequently very small chi-squares can be observed, as e.g., in the above example where it is as small as 0.0039. This means that the chi-square is too small not to reject the null-hypothesis. The results are closer to what can be expected than compatible with the assumption of a normal distribution. The obvious explanation is that Mendel somewhat mispresented his data

Phenotype	A	a
B	AB 27	aB 271
b	Ab 9	ab 93

Though Mendel may have somewhat fudged some of his data, he started a novel science that now 140 years later is the largest growing field in biomedicine. This novel science, although in its first steps, already has a major impact on the life of all of us. E.g., obtaining enough drugs, like insulin and many others, to treat illnesses worldwide was a problem that has been solved by recombinant DNA technology which enabled through genetic engineering of bacteria or yeasts the large scale production of various pharmaceutical compounds. The science of genes, often called genomics, is vast, and this chapter only briefly mentions a few statistical techniques developed for processing data of genetic research. We will start with the explanation of a few terms typically used in genomics.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 331–339.

Table 2. Bayes' Theorem, an important approach for the analysis of genetic data:example

Based on historical data the chance for girls in a particular family of being carrier for the hemophilia A gene is 50%. Those who are carrier will have a chance of ½ x ½ = ¼ = 25% that two sons are healthy. Those who are no carriers will have a 100 % chance of two healthy sons. This would mean that a girl from this population who had two healthy sons is 500 / 125 = 4 times more likely to be no carrier than to be carrier. In terms of Bayes' Theorem:
posterior odds = prior odds x likelihood ratio.
prior probability of being carrier = 50%
prior odds = 50 : 50 = 1.0
likelihood ratio = probability for carrier of having two healthy sons/ probability for non-carrier of having two healthy sons = 25% / 100% = 0.25 posterior odds = 1.0 times 0.25 = 25% or 1 in 4:
if you saw many girls from this family you would see one carrier for every 4 non-carriers.

mothers with two sons who are:	carrier	no carrier
	n = 500	n = 500
two sons healthy	n = 125	n = 500
two sons not healthy	n = 375	n = 0

2. SOME TERMINOLOGY

Bayes' Theorem (Table 2) Posterior odds = likelihood ratio x prior odds
This approach is required for making predictions from genetic data. Although the general concept of including prior evidence in the statistical analysis of clinical trial data is appealing, this concept should not be applied in usual null-hypothesis testing, because we would have to violate the main assumption of null-hypothesis testing that H0 and H1 have the same frequency distribution.

Posterior odds (Table 2) Prior odds adjusted for likelihood ratio.

Prior odds (Table 2) Prior probability of being a carrier / prior probability of being no carrier.

Likelihood ratio (Table 2) Probability for carriers of having healthy offspring/ probability for non-carrier of having healthy offspring.

Genetic linkage When 2 genes or DNA sequences are located near each other on the same chromosome, they are linked. When they are not close, crossing over occurs frequently. However, when they are close

	they tend to be inherited together. Genetic linkage is useful in genetic diagnosis and mapping because once you know that the disease gene is linked to a particular DNA sequence that is close, the latter can be used as a marker to identify the disease gene indirectly. Bayes´ Theorem can be used to combine experimental data with prior linkage probabilities as established.
Autosomal	Not x- or y-chromosome linked.
Heterosomal	X-or y-chromosome linked.
Dominant gene	Gene that is expressed in the phenotype.
Recessive gene	Gene that is expressed in the phenotype only if it is present in two complementary chromosomes.
Haplotype	Group of genetic markers linked together on a single chromosome, such as a group of DNA-sequences.
Haploid genome	Chromosomes of haploid cell (23 chromosomes, 50,000-100,000 genes).
Diploid cell	Cell with 46 chromosomes.
Chromosome	2,000-5,000 genes.
Chromosomal microband	50-100 genes.
Gene	$1,5\text{-}2000.10^3$ base-pairs.
Genomic medicine	Use of genotypic analysis to enhance quality of care.
Complex disease traits	Multifactorial diseases where multiple genes and non-genetic factors interact.
Allele	Gene derived from one parent.
Homozygous	Having identical alleles.
Heterozygous	Having different alleles.
DNA- cloning	Isolation of DNA fragments and their insertion into the nucleic acid from another biologic vector for manipulation.
DNA probe	Cloned DNA fragment used for diagnostic or therapeutic purpose.
Hybridization of single stranded DNA	Double-stranded DNA is dissociated into single-stranded, which can then be used to detect complementary strands.
Blotting procedures	Southern, Northern, Immuno-, Western blotting are all procedures to hybridize target DNA in solution to known DNA-sequences fixed on a membrane support.
Polymerase chain reaction	Oligonucleotide of known nucleic acid sequence is incubated with the target DNA and then amplified with DNA polymerase.
DNA chips	Arrays of oligonucleotides on miniature supports developed for the analysis of unknown DNA

	sequences, taking advantage of the complementary nature of nucleic acid interaction.
Mutations	Changes in DNA either heritable or obtained.
Introns	Non-coding regions of the gene.
Exons	Coding regions of the gene.
Single gene disorders	One gene plays a predominant role in determining disease.
Genotype	Chemical structure of a gene.
Phenotype	Clinical characteristics of a gene.
Gene expression	Regulation of gene function is mediated at a transcriptional level through helix-turn-helix proteins and at a posttranscriptional level through various hormones, autacoids and many more factors.

3. GENETICS, GENOMICS, PROTEONOMICS, DATA MINING

In the past two or three decades the role of genetic determinants have increased enormously in biomedical research. Of several monogenetic diseases the genetic foundation has been clarified almost completely (e.g. Huntington's disease), and of others the contribution of many genetic markers has been proved: for instance the brca 1 and 2 genes in breast cancer[1], and the mismatch gene mutations in coloncarcinoma.[2] Simultaneously, the human genome project has been the catalyst for the development of several high-throughput technologies that have made it possible to map and sequence complex genomes. These technologies are used, and will be used increasingly in clinical trials for many purposes but predominantly to identify genetic variants, and differentially expressed genes that are associated with better or worse clinical efficacy in clinical trials. In addition, the proteins associated with these genes are being investigated to disentangle their roles in the biochemical and physiological pathways of the disease and the treatment that is being studied. Together these technologies are called (high-throughput) genetics, genomics, and proteomics.

The technological advancements have made it possible to measure thousands of genes/proteins of a single patient simultaneously, and the possibility to evaluate the role of each gene/protein in differentiating between e.g. responders and non-responders to therapy. This has increased the statistical problem of multiple testing hugely, but also has stimulated research into statistical methods to deal with it. In addition methods have been developed to consider the role of clusters of genes. In this chapter we will describe a number of these new techniques for the analysis of high throughput genetic data, and for the analysis of gene-expression data. We restrict the discussion to data that are typically sampled in clinical trials including unrelated individuals only. Familial data are extremely important to investigate genetic associations: their clustered structure requires dedicated statistical techniques but these fall outside the scope of this chapter.

4. GENOMICS

In the mid-1970s, molecular biologists developed molecular cloning and DNA sequencing. Automated DNA sequencing and the invention of the polymerase chain reaction (PCR) made it possible to sequence the entire human genome. This has lead to the development of microarrays, sometimes known as DNA-chip technology. Microarrays are ordered sets of DNA molecules of known sequence. Usually rectangular, they can consist of a few hundred to thousands of sets. Each individual feature goes on the array at a precisely defined location on the substrate, and thereafter, labeled cDNA from a test and a reference RNA sample are pooled and co-hybridized. Labeling can be done in several ways, but is usually done with different fluorescently labeled nucleotides (usually Cy5-dCTP for reference, and Cy3-dCTP for test RNA). After stimulation, the expression of these genes can be measured. This involves quantifying the test and reference signals of each fluorophore for each element on the array, traditionally by confocal laser scanning. The ratio of the test and reference signals is commonly used to indicate whether genes have differential expression. Many resources are available on the web concerning the production of microarrays, and about designing microarray experiments (e.g.: 123genomics.homestead.com). A useful textbook is that of Jordan[3].

An example of a microarray is given in Figure 1. This concerns the differential expression of about 500 genes in tumour tissue of a single patient with gastric tumour.

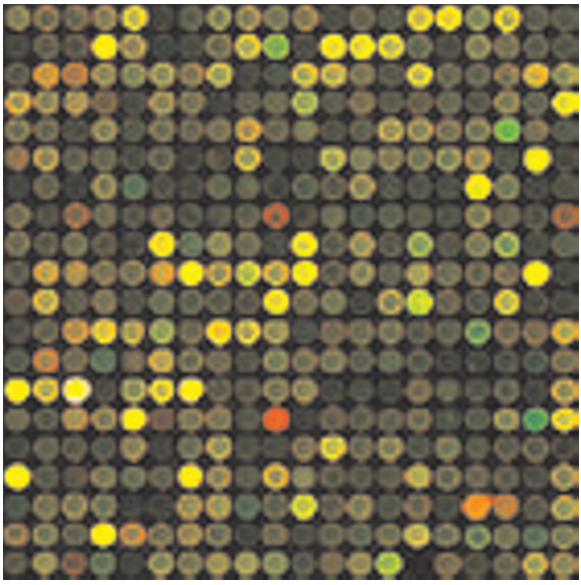


Figure 1. Example of microarray of different expression of about 500 genes in tumour tissue of a single patient.

Each spot in this chip represents a different gene, and the ratio of the two fluorescent dyes indicates whether the genes are over-expressed (dark) or under-expressed (pale) in the tumor tissue with respect to normal tissue. The transformation of the image into gene expression numbers is not trivial: the spots have to be identified on the chip, their boundaries defined, the fluorescence intensity measured, and compared to the background intensity. Usually this 'image processing' is done automatically by the image analysis software, but sometimes laborious manual adjustments are necessary. One of the most popular systems for image analysis is ScanAlyze (http://rana.stanford.edu/software).

After the image analysis, differential expression is measured by a so-called normalized ratio of the two fluorescence signals, normalized to several experimental factors. The normalized ratios of the array in Figure 1 are given in Figure 2. On the x-axis are given the 500 genes, and on the y-axis is given the normalized ratio of each gene.

It is obvious that most genes have a ratio around unity, but three or four genes are highly over-expressed with ratios above two. It is typically assumed that ratios

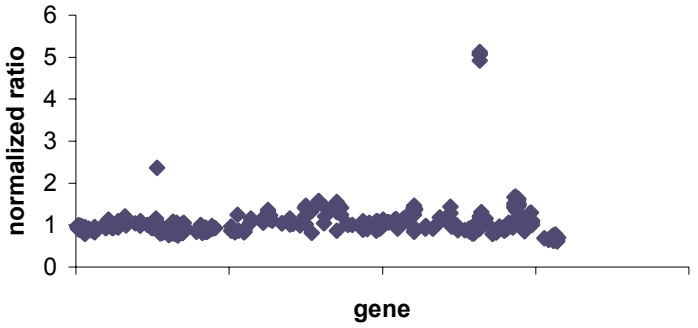


Figure 2. Normalized ratios of the array from Figure 1.

larger than 1.5 or 2.0 are indicative of a significant change in gene expression. These estimates are very crude, however, because the reliability of ratios depends on the two absolute intensities. On statistical grounds, moreover, we would expect a number of genes to show differential expression purely by chance.[4]

One way of circumventing the multiple testing problem here, is to use a mixture model.[5] Usually, it is assumed that the sample of ratios consists of subgroups of genes with normal, under-, and over-expression. In each subgroup, the ratios are mostly assumed to be normally distributed. When the sample is large enough, the percentage of normal, under-, and over-expressed genes, and associated mean ratios and standard deviations can be estimated from the data. This can be done with the logarithmically transformed ratios. The histogram of the log-transformed ratios in

Figure 2 is given in Figure 3, together with the three estimated normal distributions. In this model the probability of each gene of being over- of under-expressed can be calculated using Bayes' theorem.

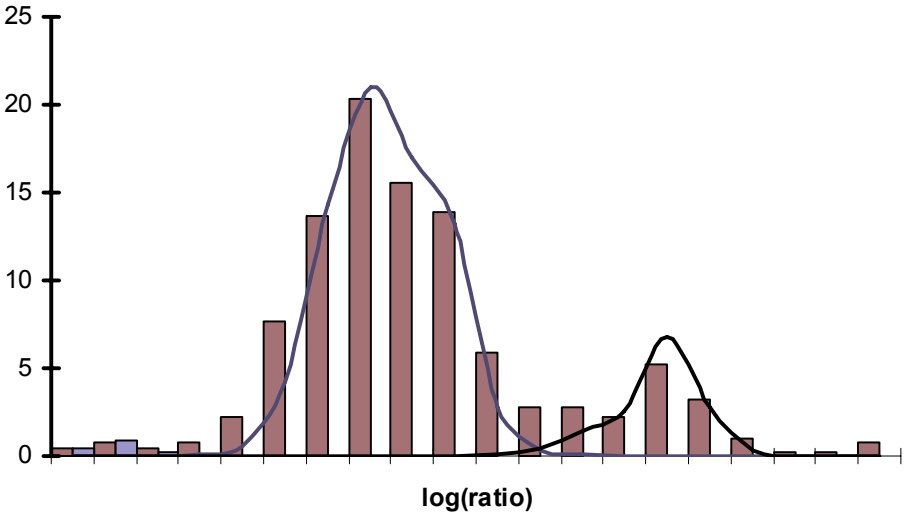


Figure 3. The histogram of the log-transformed ratios from Figure 2, calculated according to bayes' Theorem.

Although under-expressed genes could not be identified in this case, over-expressed genes were clearly seen, represented by the second mode to the right. Actually it was estimated that 14% of the genes showed over-expression, corresponding with ratios larger than 1.3.

Above is illustrated how to look at the data of a single microarray. For the analysis of a set of microarrays several different approaches are used. Two distinctions can be used: supervised or unsupervised data analysis, and hypotheses-driven or data-mining. For supervised data analysis additional data must be available to which the expression data can be related. In clinical trials a major question is often how responders and non-responders can be distinguished. Relating such response data to expression data can be done using well known techniques such as discriminant-analysis, or logistic regression. Since there may be hundreds or thousands of expression variables, one must be careful in applying these techniques, and cross-validation is often extremely useful.[6] Unsupervised data analysis is usually done by cluster analysis or principal component analysis to find groups of co-regulated genes or related samples. These techniques are often applied without specific prior knowledge on which genes are involved in which case the analysis is a kind of data-

mining. An example of a hypothesis driven analysis is to pick a potential interesting gene, and then find a group of similar or anti-correlated expression profiles.

Cluster-analysis is the most popular method currently used as the first step in gene expression analysis. Several variants have been developed: hierarchical[7], and k-means[8] clustering, self-organizing maps[9], and gene-shaving[10], and there are many more. All aim at finding groups of genes with similar properties. These techniques can be viewed as a dimensionality reduction technique, since the many thousands of genes are reduced to a few groups of similarly behaving genes. Again many tools are available on the web, and a useful site to start searching is: www.microarray.org. We used Michael Eisen's package[7] to cluster the expression data of 18 patients with gastric cancer. The typical output of a hierarchical clustering analysis is given in Figure 4. This is a dendogram illustrating the similarities between patients, a similar graph can be obtained illustrating similarities between genes. In the present case one might conclude that patients 2,6,5,7,3,13,9,10,1 and 8 form a cluster, and patients 14,15,4,11,16,17,12, and 18 another cluster. But identifying more clusters may be meaningful too.

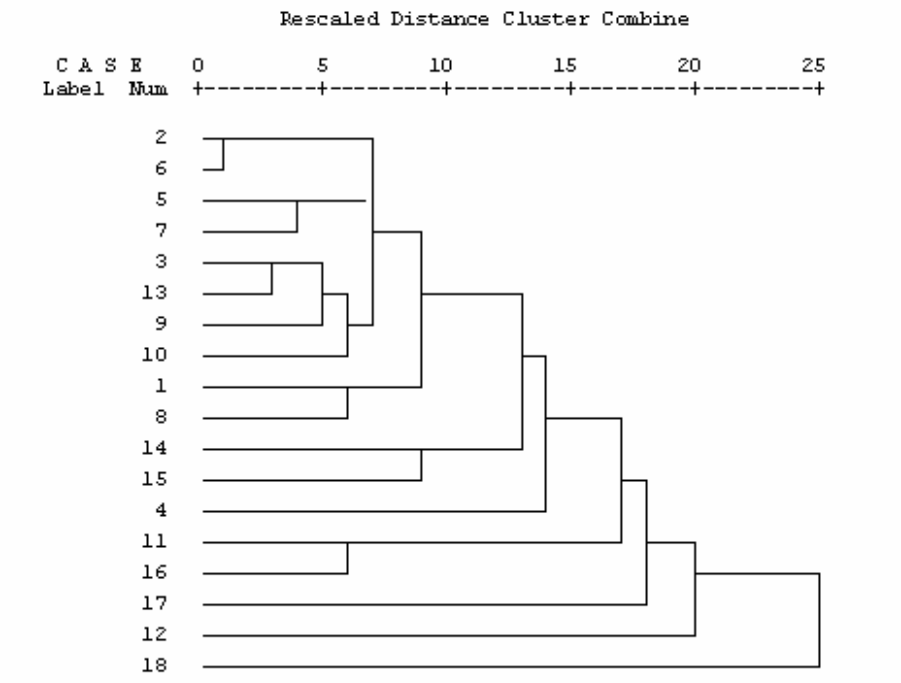


Figure 4. The typical hierarchical clustering analysis of the expression data of 18 patients with gastric cancer.

In a K-means cluster analysis the number of clusters must be specified a priori. When we specify two clusters, the same solution is found as above.

The above results illustrate that many subjective decisions need to be made in a cluster analysis, and such analysis cannot be regarded as hypothesis-driven; the primary output of a cluster analysis are new hypotheses concerning differential expressions.

5. CONCLUSIONS

Although high throughput methods are still relatively expensive, and are not used routinely in clinical trials, these methods undoubtedly will be used more often in the future. Their promise of identifying subgroups of patients with varying drug response is of major importance and is a major topic of pharmaco-genomics. In addition, differential expression profiles, and proteomics are of major importance of identifying new pathways for targeting new drugs. More sophisticated statistical methods are required, and will be developed.

6. REFERENCES

1. Cornelisse CJ, Cornelis RS, Devilee P. Genes responsible for familial breast cancer. Pathol Res Pract 1996 Jul;192(7):684-93.
2. Wijnen JT, Vasen HF, Khan PM, Zwinderman AH, van der Klift H, Mulder A, Tops C, Moller P, Fodde R. Clinical findings with implications for genetic testing in families with clustering of colorectal cancer. N Engl J Med 1998 Aug 20;339(8):511-8 .
3. Jordan B (Ed.). DNA Microarrays: gene expression applications. Berlin: Springer-Verlag, 2001.
4. Claverie JM. Computational methods for the identification of differential and coordinated gene expression. Hum Mol Genet 2001; 8 (10): 1821-32.
5. McLachlan G. Mixture.model clustering of microarray expression data. Aus Biometrics and New Zealand Stat Association Joint Conference, 2001, Christchurch, New Zealand.
6. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000, 403: 503-11.
7. Eisen M et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 1998, 95: 14863-14867.
8. Tavazoie et al. Systematic determination of genetic network architecture. Nat Genet 1999, 22: 281-5.
9. Tamayo et al. Interpreting patterns of gene-expression with self-organizing maps. Proc Natl Acad Sci USA, 1999, 96: 2907-12.
10. Tibshirani et al. Clustering methods for the analysis of DNA microarray data. Tech. rep. Stanford University, Dept of Statistics, Stanford.

CHAPTER 28

RELATIONSHIP AMONG STATISTICAL DISTRIBUTIONS

1. INTRODUCTION

Samples of clinical data are frequently assessed through 3 variables:

The mean result of the data.
The spread or variability of the data.
The sample size.

Generally, we are primarily interested in the first variable, but mean or proportion does not tell the whole story, and the spread of the data may be more relevant. For example, when studying how two drugs reach various organs, the mean level may be the same for both, but one drug may be more variable than the other. In some cases, too little and, in other cases, dangerously high levels get through. The Chi-square-distribution, unlike the normal distribution, is used for the assessment of such variabilities. Clinical scientists although they are generally familiar with the concept of null-hypothesis-testing of normally distributed data, have difficulties to understand the null-hypothesis testing of Chi-square-distributed data, and do not know how closely Chi-square is related to the normal-distribution or the T-distribution. The Chi-square-distribution has a relatively young history. It has been invented by K. Pearson[1] one hundred years ago, three hundred years after the invention of the normal-distribution (A. de Moivre 1667-1754). The Chi-square-distribution and its extensions have become the basis of modern statistics and have provided statisticians with a relatively simple device to analyze complex data, including multiple groups and multivariable variables analyses. The present paper was written for clinical investigators / scientists in order to better understand the relation between normal and chi-square distribution, and how they are being applied for the purpose of null-hypothesis testing.

2. VARIANCES

Repeated observations exhibit a central tendency, the mean, but, in addition, exhibit spread or dispersion, the tendency to depart from central tendency. If measurement of central tendency is thought of as good bets, then measures of spread represent the poorness of central tendency otherwise called deviation or error. The larger such deviations are, the more do cases differ from each other and the more spread does

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 341–354.

the distribution show. What we need is an index to reflect this spread or variability. First of all, why not simply take the average of the deviations (d-values) about the mean as measure of variability:

$\Sigma (d / n)$ where n = sample size.

This, however, will not work, because when we add up negative and positive departures from the mean, our overall variance will equal zero. A device to get around this difficulty is to take the square of each deviation:

$\Sigma (d^2 / n)$ is defined the variance of n observations.

$\Sigma (d / n)$, although it can not be used as index to reflex variability, can be readily used to define the mean of a sample of repeated observations, if the size of observations is taken as distance from zero rather than mean. Suddenly, means and variances look a lot the same, and it is no surprise that statistical curves and tables used to assess either of them are closely related. A Chi-square-distribution is nothing else than the distribution of square values of a normal-distribution. Null-hypothesis-testing-of-variances is much similar to null-hypothesis-testing-of-means. With the latter we reject the null-hypothesis of no effect if our mean is more than 1.96 SEMs (standard errors of the mean) distant from zero. With the latter we reject the null-hypothesis of no effect if our standardized variance is more than "1.96 ^{2} SEMs^2" distant from zero. Because variances are squared and, thus, non-negative values, the Chi-square approach can be extended to test hypotheses about many samples. When variances or add-up variances of many samples are larger than allowed for by the Chi-square-distribution-graphs, we reject the probability that our results are from normal distributions, and conclude that our results are significantly different from zero. The Chi-square test is not only adequate to test multiple samples simultaneously, but is also the basis of analysis of variance (ANOVA).

3. THE NORMAL DISTRIBUTION

The normal distribution curve can be drawn from the formula below.

$$f(x) = \frac{1}{\sqrt{2\pi s^2}} e^{-(x-m)^2/2s^2}$$

where s = standard deviation and m = mean value.

Repeated observations in nature do not precisely follow this single mathematical formula, and may even follow largely different patterns. The formula is just an approximation. And so, it is remarkable that the approach works in practice, although the p-values obtained from it are sometimes given inappropriate emphasis. We should not forget that a p-value of <0.001 does not mean that we have proven something for the entire population, but rather that we have proven something on the understanding that our data follow a normal distribution and that our data are representative for the entire population. Frequently, the results as provided by clinical trials are much better than those observed in general practice, because the population follows a different frequency distribution or because the enrollees in a trial are selected groups not representative for the entire population. We wish that more often these possibilities would be accounted by the advocates of evidence-

based medicine. If we are willing to accept the above limitations, the normal distribution can be used to try and make predictions, with the understanding that statistical testing cannot give certainties, only chances. How was the normal distribution invented? At first, investigators described their data in the form of histograms (figure 1 upper graph: on the x-axis the individual data and on the y-axis how often).

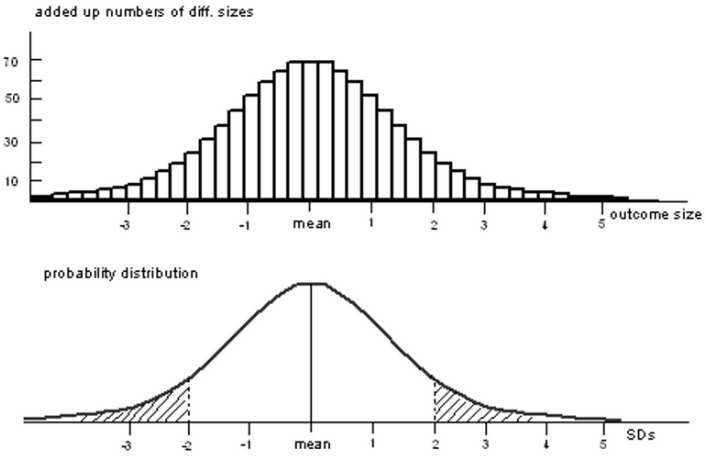


Figure 1. *Upper graph shows histogram: on the x-axis we have the individual data and on the y-axis we have "how often" (the mean value is observed most frequently, while the bars on both side of the mean gradually grew shorter). Lower graph shows normal distribution: the bars on the y-axis have been replaced with a continuous line, it is now impossible to read from the graph how many patients had a particular outcome. Instead, we infer that the total area under the curve (AUC) represents 100% of our data, AUC left from the mean represents 50%, left from – 1 SD (standard deviation) approximately 15% of the data, and left from -2 SDs approximately 2.5 % of the data. This curve although suitable for describing a sample of repeated observations, is not yet adequate for testing statistical hypotheses.*

Often, the mean value is observed most frequently, while the bars on both side of the mean gradually grow shorter. From this histogram to a normal distribution curve is a short step (Figure 1 lower graph). The bars on the y-axis have been replaced by a continuous line. It is now impossible to read from the graph how many patients had a particular outcome. Instead, relevant inferences can be made: the total area under the curve (AUC) presents 100% of our data, AUC left from the mean presents 50%, left from - 1SD (standard deviation) approximately 15% of the data, and left from -2 SDs approximately 2.5 % of the data. This curve although suitable for describing a sample of repeated observations, is not yet adequate for testing statistical hypotheses. For that purpose, a narrow normal curve is required (Figure 2).

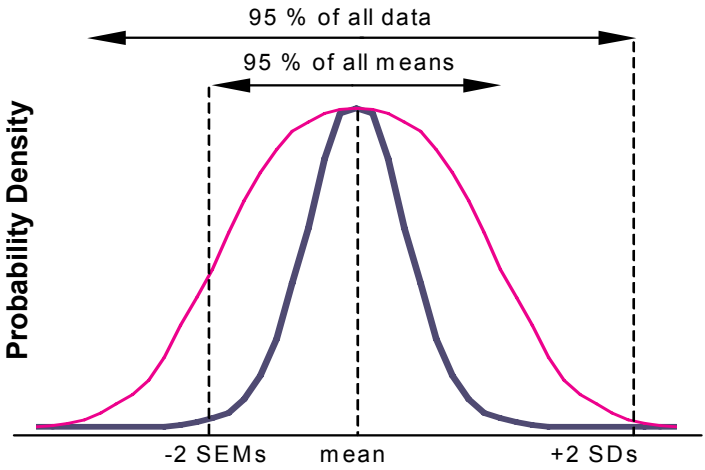


Figure 2. *Narrow and wide normal curve: the wide one summarizes the data of our trial, the narrow one summarizes the means of many trials similar to our trial.*

The narrow and wide curve from Figure 2 are both based on the same data, but have different meaning. The wide (with SDs on the x-axis) one summarizes the data of our trial, the narrow one (with SEMs (standard errors of the mean) on the x-axis) summarizes the means of many trials similar to ours. This may be difficult to understand, but our sample is representative, and it is easy to conceive that the distributions of means of many similar samples from the same population will be narrower and have fewer outliers than the distribution of the actual data. This concept is relevant, because we want to use it for making predictions from our data to the entire population.

We should add here that there is only a small difference between the normal and the t - distribution. The latter is a bit wider with small numbers. The chi-square distribution makes no difference between normally and t - like distributed data.

4. NULL-HYPOTHESIS TESTING WITH THE NORMAL OR T-DISTRIBUTION

What does “null-hypothesis” mean: we hypothesize that if the result of our trial is not different from zero, we have a negative trial. What does the null-hypothesis look like in graph? Figure 3 shows H1, the graph based on the data of our trial with SEMs on the x-axis (z-axis), and H0, the same graph with a mean of 0.

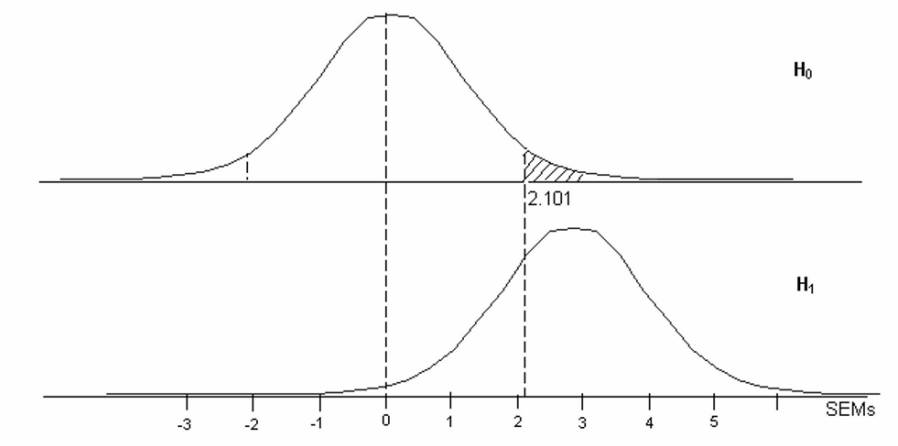


Figure 3. *H1 is the graph based on the data of our trial with SEMs on the x-axis (z-axis), and H0 is the same graph with mean 0 (mean ± SEM = 0 ± 1).*

Now we make a giant leap from our data to the entire population, and we can do so, because we assume, that our data are representative for the entire population. H1 is also the summary of the means of many trials similar to our trial. If we repeated the trial, differences would be small and the summary would look alike. H0 is also the summary of the means of many trials similar to our trial, but with an overall effect of 0. Our mean is not 0, but 2.9. Still it could be an outlier of may studies with an overall effect of 0. So, think from now on of H0 as distribution of the means of many trials with overall effect 0. If hypothesis 0 is true, then the mean of our study is part of H0. We can not prove this, but we can calculate the chance/probability of this possibility. A mean result of 2.9 is far distant from 0. Suppose it belongs to H0. Only 5% of the H0-trials are more than 2.1 SEMs distant from 0, because the AUC of H0 = 5%. Thus, the chance that it belongs to H0 is less than 5%. We reject the null-hypothesis of no effect concluding that there is less than 5% chance to find this result. In usual terms, we reject the null-hypothesis of no effect at $p < 0.05$ or < 5%.

5. RELATIONSHIP BETWEEN THE NORMAL-DISTRIBUTION AND CHI-SQUARE-DISTRIBUTION, NULL-HYPOTHESIS TESTING WITH CHI-SQUARE DISTRIBUTION

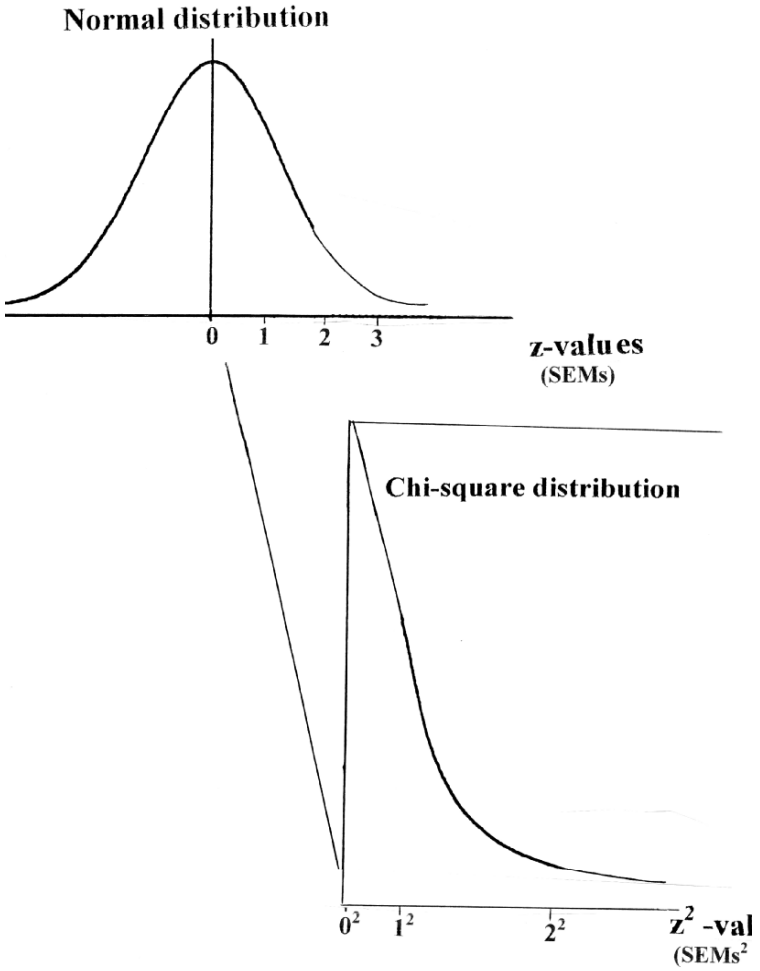


Figure 4. *Upper graph shows a normal distribution. Lower graph shows what happens if the x-values of this normal-like-curve are squared. The normal-curve changes into a Chi-square-curve.*

The upper graph of Figure 4 shows a normal distribution, on the x-axis individual data expressed as distances from the mean, and on the y-axis "how often" the individual data are being observed. The lower graph of Figure 4 shows what happens of the x-values of this normal distribution is squared. We get no negative x-values anymore, and the x-values 0 and 1 give rise to y-values twice the size, while the new curve is skewed to the right: the new curve is what we call a chi-

square curve. The upper curve is used to test the null-hypothesis that the mean result of our trial is significantly different from zero, the lower one to test that our variance is significantly different from zero.

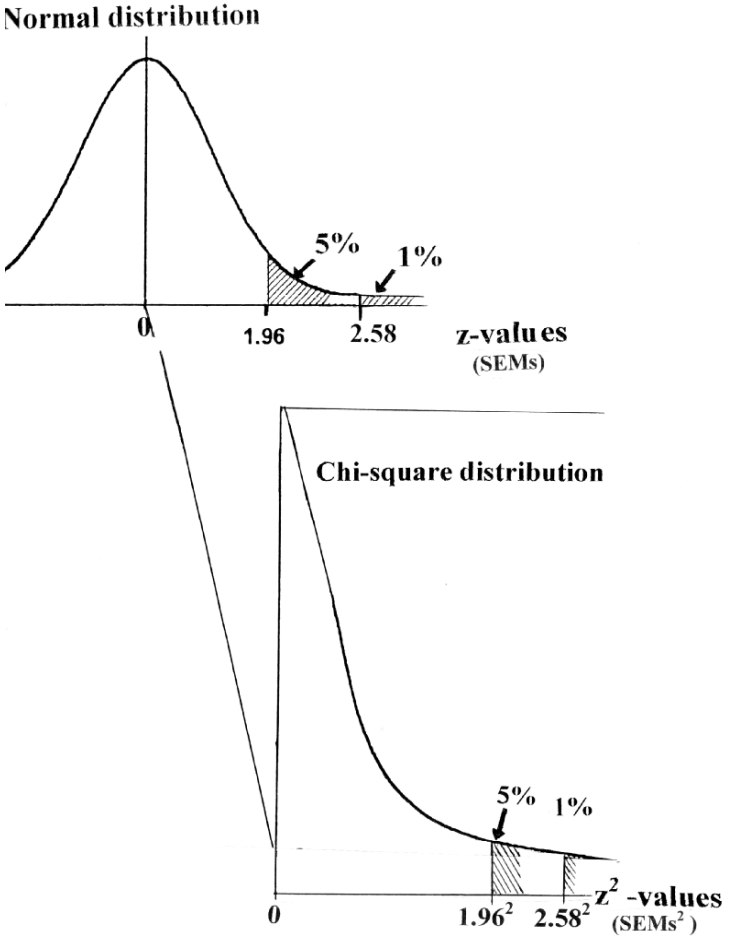


Figure 5. *Upper graph gives the x-values, otherwise called z - values, of a null-hypothesis of a real normal-distribution. Lower graph shows what happens when z - values are squared. The z - distribution turns into a non-negative Chi-square- distribution. Upper graph: with z > 1.96 the right-end AUC < 5%; lower graph: with* $z^2 > (1.96)^2$ *the right-end AUC < 5%.*

Figure 5 shows how things work in practice. The upper graph gives on the x-axis the possible mean result or our trial expressed in units of SEMs, otherwise called z-value, or, with t - test, t - value. On the y-axis we "how often this result will be obtained". If our mean result is more than approximately 2 SEMs (or with normal distribution precisely 1.96 SEMs) distant from zero, this will happen in 5% of the

cases, because the AUC right from 1.96 SEMs is 5%. If more than 2.58 distant from zero, this will happen in 1% of the cases. With a result that far from zero we reject the null-hypothesis that our result is not different from 0, and conclude that, obviously, our data are significantly different from 0, at p<5% or 1% (<0.05 or <0.01).

Figure 5 lower graph gives a draft of the possible variances of our trial. On the x-axis we have the variance of our trial expressed in units $(SEMs)^2$, otherwise called z^2- values. On the y-axis we have again "how often this variance will be obtained". For example, if our variance is more than 1.96^2 $SEMs^2$ distant from zero, this will happen in less than 5% of the cases. This is so, because the AUC right from $z^2 = 1.96^2$ is 5% of the total AUC of 100%. If our variance is more than $z^2 = 2.58^2$ distant from zero, this chance is 1%. We reject the null-hypothesis that our variance is not significantly different from 0 and we do so at a probability of 1% ($p < 0.01$).

6. EXAMPLES OF DATA WHERE VARIANCE IS MORE IMPORTANT THAN MEAN

The effects on circadian glucose levels of slow-release-insulin and acute-release-insulin are different. The mean glucose-level is the same for both treatment formulas, but the latter formula produces more low and high glucose levels. Spread or variance of the data is a more important determinant of treatment effect than is the mean glucose value.

A pill producing device is approved only if it will produce pills with a SD not larger than e.g. 6mg. Rather than mean the variance of a test-sample is required to test the device.

People on selective serotonin reuptake inhibitors (SSRIs) may not only show a lower average of performance, but also a higher variability in performance relative to their counterparts. Variance, in addition to average of performance is required to allow for predictions on performances.

The variability in stay-days in hospital is more relevant than the mean stay-days, because greater variability is accompanied with a more demanding type of care.

Why should we statistically test such questions anyway? Or why not simply calculate the mean result and standard deviation of a sample of data, and, then, check if the SD is within a predefined area. We, subsequently, accept this as sufficient probability to make further predictions about future observations. However, by doing so we will never know the size of this probability. A statistical test rejects the null-hypothesis of no difference from 0 at a 5 % or lower level of probability, and this procedure is widely valued as a powerful aid to erroneous conclusions. A more extensive overview of current routine methods to assess variability of data samples is given in chapter 26.

7. CHI-SQUARE CAN BE USED FOR MULTIPLE SAMPLES OF DATA

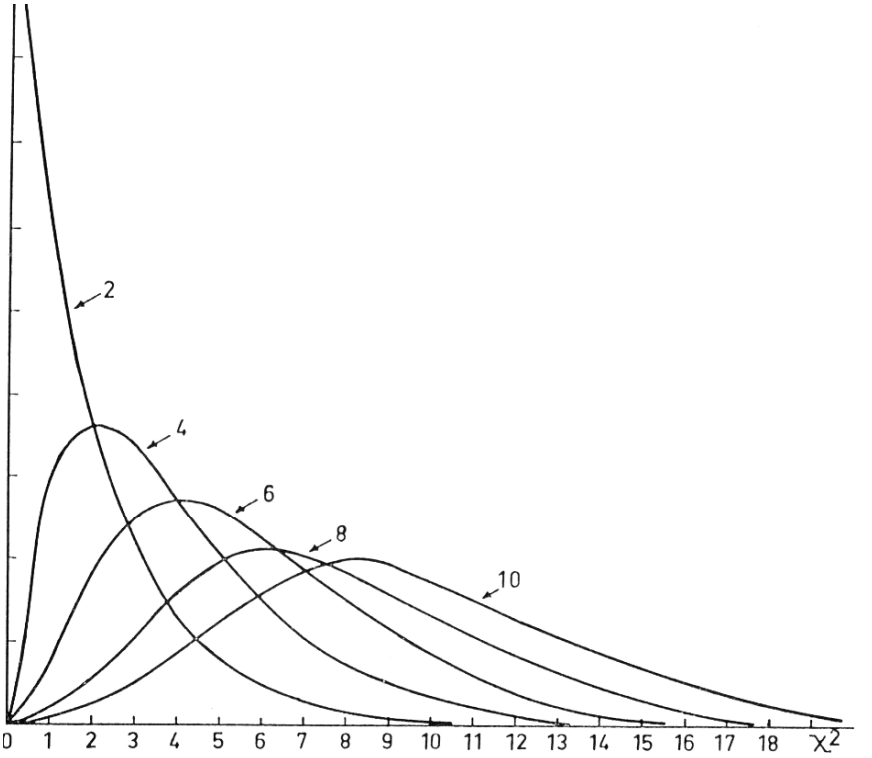


Figure 6. *The general form of the Chi-square distributions for larger samples of data.*

1. Contingency tables

The simplest extension of the chi-square test is the analysis of a two-by-two contingency table. With contingency tables we want to test whether two groups of binary data (yes/no data) are significantly different from one another. We have 4 cells ((1) group-1 yes, (2) group-1 no, (3) group-2 yes, (4) group-2 no). The null-hypothesis is tested by adding up:

$$\text{chi-square} = \frac{(O-E)_{\text{cell 1}}^{\ 2}}{E_{\text{cell 1}}} + \frac{(O-E)_{\text{cell 2}}^{\ 2}}{E_{\text{cell 2}}} + \frac{(O-E)_{\text{cell 3}}^{\ 2}}{E_{\text{cell 3}}} + \frac{(O-E)_{\text{cell 4}}^{\ 2}}{E_{\text{cell 4}}}$$

where O means observed numbers, and E means expected numbers per cell if no difference between the two groups is true (the null-hypothesis). The E-value in the denominator standardizes the test-statistic.

2. Pooling relative risks or odds ratios in a meta-analysis of multiple trials

In meta-analyses the results of the individual trials are pooled in order to provide a more powerful assessment. Chi-square-statistic is adequate for testing a pooled result. The natural logarithms are used to approximate normality.

$$\text{Chi-square} = \frac{\left(\frac{\ln RR_1}{s_1^2} + \frac{\ln RR_2}{s_2^2} + \frac{\ln RR_3}{s_3^2} \ldots \right)^2}{\frac{1}{s_1^2} + \frac{1}{s_2^2} + \frac{1}{s_3^2} + \ldots}$$

where RR means relative risk and s means SD of this relative risk per sample. The $1/s^2$ -term in the denominator again standardizes the test-statistic.

3. Analysis of variance (ANOVA)

Unlike the normal-test or the t - test, the Chi-square-test can be extended to testing more than one sample of data simultaneously. Variances are non-negative values, and they can simply be added up. This is, actually, the way variance is defined, the add-up sum of squared distances from the mean. Any subsequent sample of data, if from a normal distribution or t - distribution can be simply added up to the first sample and the add-up sum can be analyzed simultaneously. And, so, with little more effort than demonstrated for 1 sample of data, multiple samples can be added to the model in order to test the null-hypothesis of no difference from zero. This is possible both for samples of continuous data and proportional data, including percentages, proportions, odds ratios, risk ratios etc. The only difference is the breath of the chi-square curve: it gets wider and wider the more samples or the more proportions we add (Figure 6).

A further extension of the use of the Chi-square-statistic is ANOVA. ANOVA makes use of the division-sum of two Chi-square-distributions. This division-sum, indeed, looks much like a usual Chi-square-distribution , as shown for example in Figure 7.

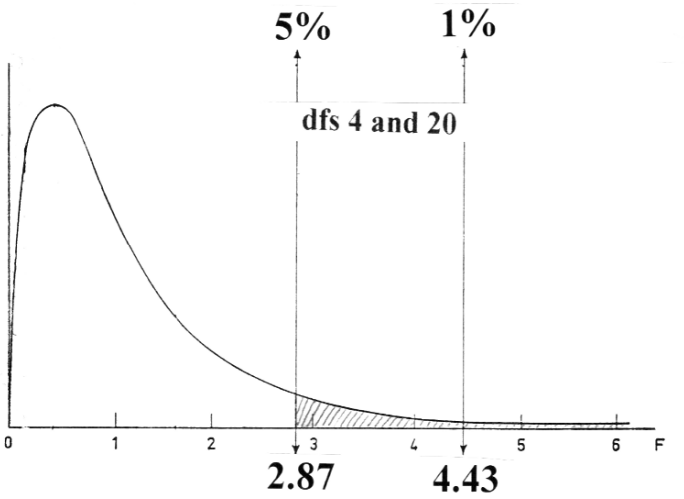


Figure 7. *Example of an F-distribution making use of the division-sum of two Chi-square-distributions with 4 and 20 degrees of freedom (dfs).*

For example, ANOVA with k groups works as follows:

Total variation

| |

Between-group-variation within-group-variation

Variations are expressed as sums of squares (SS) and can be added up to obtain the total variation.

We assess whether between-group-variation is large compared to within-group-variation.

Group	n patients	mean	sd
1	-	-	-
2	-	-	-
3	-	-	-
...			
k			

Grand mean = (mean 1 + 2 +3+..k) / k

$SS_{between\ groups} = n_1\ (mean_1 - grand\ mean)^2 + n_2\ (mean_2 - grand\ mean)^2 + \ldots.$

$SS_{within\ groups} = (n_1\text{-}1)\ (sd_1^{\ 2}) + (n_2\text{-}1)\ (sd_2^{\ 2}) + \ldots..$

$$F = test\text{-}statistic = \frac{SS_{between\ groups} / dfs^*}{SS_{within\ groups} / dfs}$$

* dfs means degrees of freedom (for SS between groups dfs = k-1, for SS within groups dfs = $n_1 + n_2 + n_3 + ..n_k$ - k).

The F-table gives p - value.

8. DISCUSSION

The current chapter is not a cook-book-like instruction for the use of various statistical methods. It only briefly examines the connection between the Chi-square-distribution and other important statistical distributions. They form the basis of all statistical inferences, which are given so much emphasis in today's clinical medicine. The Chi-square-distribution is directly derived from the normal-distribution. The F-distribution is directly derived from the Chi-square distribution. Over and over again, these distributions have shown their utility in the solution of problems in statistical inference. However, none of these distributions is empirical in the sense that someone has taken a large number of samples and found that the sample values actually follow the same mathematical function. Of course, nature does not follow a single mathematical function. The function is an approximation, but it performs well and has proven to be helpful in making clinical predictions. The distribution is also based on assumptions, and, like other theory-based assessments, deals with "if-then" statements. That is why the assumptions about representative samples and normal-distribution in our sample are so important. If we apply the theory of statistics for making inferences from samples, we cannot expect this theory to provide us with adequate answers unless conditions specified in the theory hold true.

Apart from the general requirement of random sampling of independent observations, the most usual assumption made is that the population-distribution is normal. The Chi-square, the t-, and the F-distributions all rest upon this assumption. The normal-distribution can be considered the "parent" distribution to the others. Similarly, there are close connections between the F-distribution and both the normal- and the Chi-square-distributions. Basically, the F-statistic is the ratio of two independent Chi-square-statistics, each of which characterized by its own degrees of freedom. Since a Chi-square-statistic is defined in terms of a normal-distribution, the F-distribution also rests upon the same assumptions, albeit of two (or more than two) normal-distributions. The Chi-square-distribution focused on in this paper is, thus, just another approach of the bell-shape-like normal distribution and is also the basic element of the F-distribution. Having some idea of the interrelations of these distributions will be of help in understanding how the Chi-square is used to test a hypothesis-of-variance, and how the F-distribution is used to test a hypothesis-about-several-variances.

We conclude that the Chi-square-distribution and its extensions have become the basis of modern statistics and have provided clinical scientists with a relatively simple device to analyze complex data, including multiple groups / multiple variances. The present chapter was written for clinical investigators/ scientists in order to better understand benefits and limitations of Chi-square-statistic and its many extensions for the analysis of experimental clinical data.

9. CONCLUSIONS

Statistical analyses of clinical data are increasingly complex. They often involve multiple groups and measures. Such data can not be assessed simply by differences between means but rather by comparing variances. The objective of this chapter was to focus on the Chi-square (χ^2)-test as a method to assess variances and test differences between variances. To give examples of clinical data where the emphasis is on variance. To assess interrelation between Chi-square and other statistical methods like normal-test (Z-test), T-test and Analysis-Of-Variance (ANOVA).

A Chi-square-distribution is nothing else than the distribution of square values of a normal-distribution. Null-hypothesis-testing-of-variances is much similar to null-hypothesis-testing-of-means. With the latter we reject the null-hypothesis of no effect if our mean is more than 1.96 SEMs (standard errors of the mean) distant from zero. With the latter we reject the null-hypothesis of no effect if our standardized variance is more than 1.96^2 $SEMs^2$ distant from zero. Because variances are squared and, thus, non-negative values, the Chi-square approach can be extended to test hypotheses about many samples. When variances or add-up variances of many samples are larger than allowed for by the Chi-square-distribution-graphs, we reject the probability that our results are from normal distributions, and conclude that our results are significantly different from zero. The Chi-square test is not only adequate to test multiple samples simultaneously, but is also the basis of ANOVA.

We conclude that the Chi-square-distribution focused on in this paper is just another approach of the bell-shape-like normal-distribution and is also the basic element of the F-distribution as used in ANOVA. Having some idea about interrelations between these distributions will be of help in understanding benefits and limitations of Chi-square-statistic and its many extensions for the analysis of experimental clinical data.

10. REFERENCES

1. Pearson K. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it cannot be reasonably supposed to have arisen from random sampling. Philosophical Magazine 1900; 50: 339-57.

CHAPTER 29

TESTING CLINICAL TRIALS FOR RANDOMNESS

1. INTRODUCTION

As it comes to well-balanced random sampling of representative experimental data, nature will be helpful to provide researchers with results that comply with the random property. It means that such data closely follow statistical frequency distributions. We continually make use of these statistical frequency distributions for analyzing the data and making predictions from them. However, we, virtually, never assess how close to the expected frequency distributions the data actually are. Unrandomness of the data may be one of the reasons for the lack of homogeneity in current research, and may jeopardize the scientific validity of research data.[1-3] Statistical tests used for the analysis of clinical trials assume that the observations represent a sample drawn at random from a target population. It means that any member of the population is as likely to be selected for the sampled group as the other. An objective procedure is required to achieve randomization. When other criteria are used to permit investigators to influence the selection of subjects, one can no longer conclude that the observed effects are due to the treatment rather than biases introduced by the process of selection. Also, when the randomization assumption is not satisfied, the logic underlying the distributions of the test statistics used to estimate that the observed effects are due to chance rather than treatment effect fails, and the resulting p-values are meaningless. Important causes for unrandomness in clinical trials include extreme exclusion criteria[4] and inappropriate data cleaning.[1]

In the present chapter we review some methods to assess clinical data for their compliance with the random property.

2. INDIVIDUAL DATA AVAILABLE

If the individual data from a clinical trial are available, there are two methods to assess the data for their compliance with the random property, the chi-square goodness of fit and the Kolmogorov-Smirnov goodness of fit tests. Both tests are based on the assumption that differences between observed and expected experimental data follow normal distributions. The two tests raise similar results. If both of them are positive, the presence of unrandomness in the data can be assumed with confidence, and efficacy analysis of the data will be a problem. The tests can be used with any kind of random data like continuous data, proportions or frequencies. In this section two examples of continuous data will be given, in the

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 355–366.

next section an example of frequencies will be given. Also, we briefly address randomness of survival data, which are increasingly used as primary endpoint variable, e.g., of the 2003 volume 362 of the Lancet in 48% of the randomized trials published.

1. Method 1: the chi-square goodness of fit test

In random populations body-weights follow a normal distribution. Is this also true for the body-weights of a group of patients treated with a weight reducing compound? The example is modified from Levin and Rubin with permission from the editor.[5]

Individual weight (kgs)
85 57 60 81 89 63 52 65 77 64
89 86 90 60 57 61 95 78 66 92
50 56 95 60 82 55 61 81 61 53
63 75 50 98 63 77 50 62 79 69
76 66 97 67 54 93 70 80 67 73

The area under the curve (AUC) of a normal distribution curve is divided into 5 equiprobable intervals of 20 % each, we expect approximately 10 patients per interval. From the data a mean and standard deviation (sd) of 71 and 15 kg are calculated. Figure 1 shows that the standardized cut-off results (z-values) for the 5 intervals are –0.84, -.025, 0.25 and 0.84. The real cut-off results are calculated according to

$$z = \text{standardized result} = \frac{\text{unstandardized result - mean result}}{\text{sd}}$$

and are given below (pts = patients).

Intervals (kgs)	$-\infty$	-	58.40	-	67.25	-	74.25	-	83.60	-	∞
As they are equiprobable, we expect per interval:	10 pts	|	10 pts	|	10 pts	|	10 pts	|	10 pts	
We do, however, observe the following numbers:	10 pts	|	16 pts	|	3 pts	|	10 pts	|	11 pts	

The chi-square value is calculated according to

$$\sum \frac{(\text{observed number-expected number})^2}{\text{expected number}} = 8.6$$

This chi-square value means that for the given degrees of freedom of 5-1 = 4 (there are 5 different intervals) the null-hypothesis of no-difference-between-observed-

and-expected can not be rejected. However, our p-value is < 0.10, and, so, there is a trend of a difference. The data may not be entirely random.

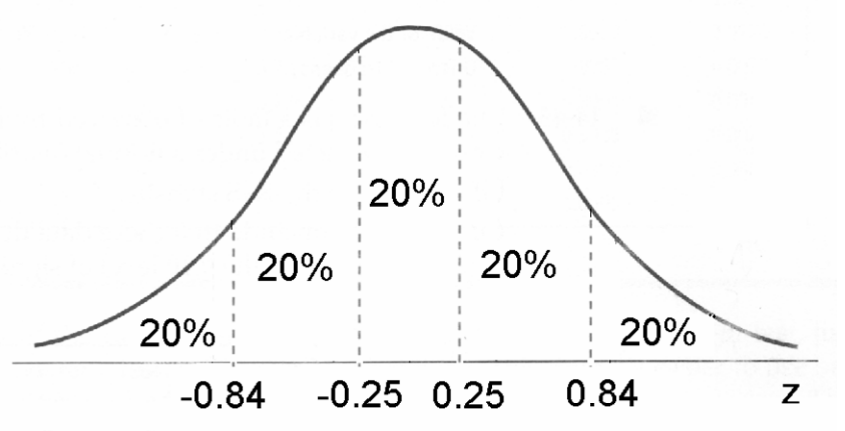


Figure 1. The standardized cut-off results (z-values) for the 5 intervals with an AUC of 20% are -0.84, -0.25, 0.25, and 0.84 (AUC = area under the curve).

2. Method 2: the Kolmogorov-Smirnov goodness of fit test

In random populations plasma cholesterol levels follow a normal distribution. Is this also true for the plasma cholesterol levels of the underneath patients treated with a cholesterol reducing compound? This example is also modified from Levin and Rubin with permission from the editor.[5]

Cholesterol (mmol/l)	<4.01	4.01-5.87	5.87-7.73	7.73-9.59	>9.59
Numbers of pts	13	158	437	122	20

The cut-off results for the 5 intervals must be standardized to find the expected normal distribution for these data according to

$$z = \text{standardized cut-off result} = \frac{\text{unstandardized result} - \text{mean result}}{\text{sd.}}$$

With a calculated mean (sd) of 6.80 (1.24) we find -2.25, -0.75, 0.75 and 2.25. Figure 2 gives the distribution graph plus AUCs. With 750 cholesterol-values in total the expected frequencies of cholesterol-values in the subsequent intervals are

12.2 x 750 = 9.2
21.4 x 750 = 160.8
54.7 x 750 = 410.1
21.4 x 750 = 160.8
12.2 x 750 = 9.2

The observed and expected frequencies are, then, listed cumulatively (cumul = cumulative) :

Frequency observed	cumul	relative (cumul/750)	expected	cumul	relative (cumul/750)	cumul observed-expected
13	13	0.0173	9.1	9.1	0.0122	0.0051
158	171	0.2280	160.9	170.0	0.2266	0.0014
437	608	0.8107	410.1	580.1	0.7734	0.0373
122	730	0.9733	160.8	740.9	0.9878	0.0145
20	750	1.000	9.1	750	1.000	0.0000

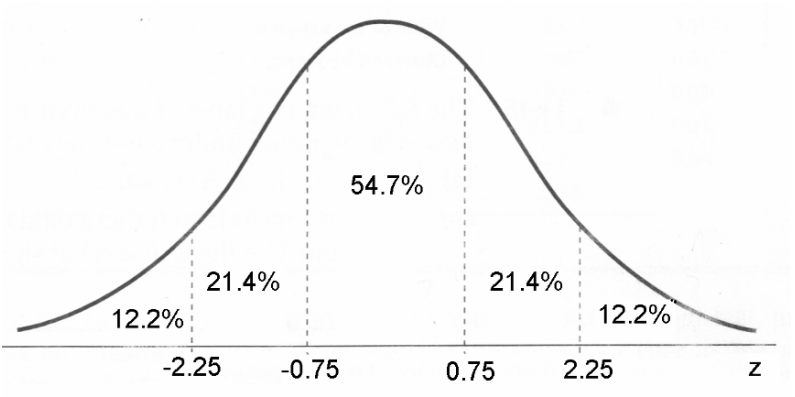


Figure 2. The standardized cut-off results (z-values) for the 5 intervals are calculated to be -2.25, -0.75, 0.75, and 2.25. Corresponding AUCs are given in the graph (AUC = area under the curve).

According to the Kolmogorov-Smirnov table (t able 1) the largest cumulative difference between observed and expected should be smaller than $1.36 / \sqrt{n} = 1.36 / \sqrt{750} = 0.0497$, while we find 0.0373. This means that these data are well normally distributed.

Table 1. Critical values of the Kolmogorov-Smirnov goodness of fit test

Sample size (n)	Level of statistical significance for maximum difference between cumulative observed and expected frequency				
n	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.463
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
Over 35	1.07 / √n	1.14 / √n	1.22 / √n	1.36 / √n	1.63 / √n

3. Randomness of survival data

Cox regression is routinely used for the analysis of survival data. It assumes that randomly sampled human beings survive according to an exponential pattern. The presence of an exponential pattern can be confirmed by logarithmic transformation. If the transformed data are significantly different from a line, the exponential relationship can be rejected. Figure 3 shows the survivals of 240 patients with small cell carcinomas, and figure 4 shows the natural logarithms of these survivals. From

figure 4 it can be observed that logarithmic transformation of the numbers of patients alive readily produces a close to linear pattern. A Pearson's correlation coefficient of these data at $p < 0.0001$ confirms that these data are closer to a line than could happen by chance. We can conclude that these survival data are compatible with a sample drawn at random.

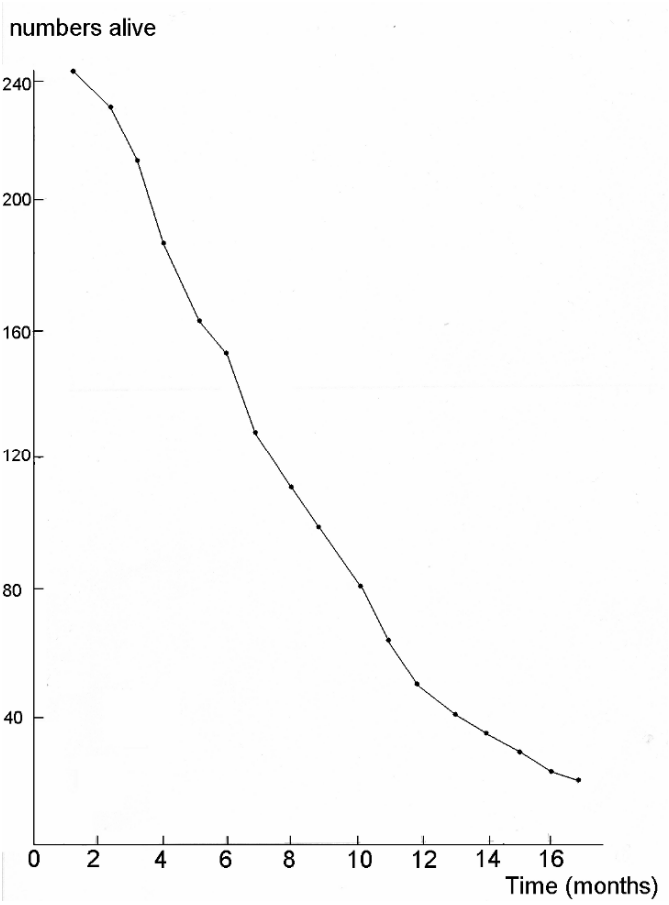


Figure 3. Survivals of 240 random patients with small cell carcinomas.

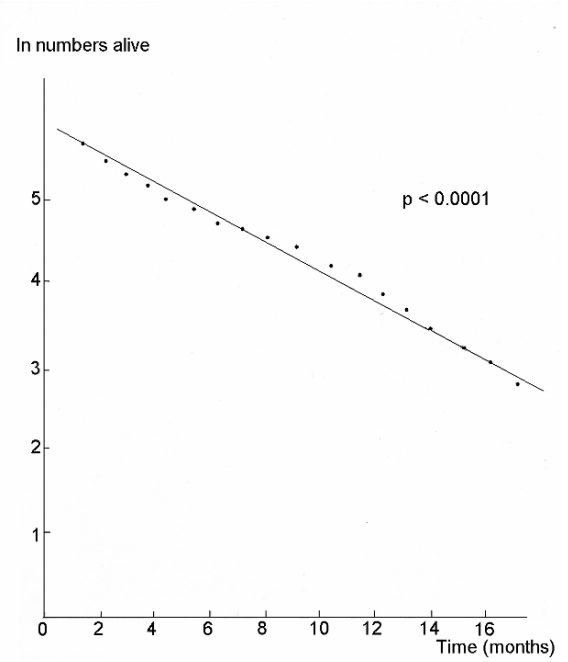


Figure 4. The logarithmic transformation of the numbers of patients from figure 3 produces a close to linear pattern.

3. INDIVIDUAL DATA NOT AVAILABLE

1. Studies with single endpoints

If the actual data from the research are not available like in most clinical reports, it is harder to assess randomness of the data. However, it is not impossible to do so. Some criteria for assessing main endpoint results of published studies for such purpose have been recently proposed by us[1,2] , and have already been addressed in chapter 10.

1. An observed p-value of < 0.0001 in a clinical trial.
   In statistics, a generally accepted concept is "the smaller the p-value, the better reliable the results". This is not entirely true with current randomized controlled trials. First, randomized controlled trials are designed to test small differences. A randomized controlled trial with major differences between old and new treatment is unethical because half of the patients have been given an inferior treatment. Second, they are designed to confirm prior evidence. For that purpose, their sample size is carefully calculated. Not only too small, but also too large a sample size is considered unethical and unscientific, because negative studies have to be repeated and a potentially inferior treatment should not be given to too many patients. Often in a study the statistical power is set at

80%. An expected power of 80% means a < 10 per cent chance of a p-value < 0.0001 with normally distributed data[6] and a < 5 per cent chance of a p-value < 0.0001 with t-distributed data and samples sizes under 50, (as often observed in, e.g., oncology trials).

2. An observed p-value of > 95 % in a clinical trial.
   P-values are generally used as a cut-off levels to indicate the chance of a difference from H_0 (the null-hypothesis-of-no-effect) in our data. The larger the p-value the smaller the chance of a difference from H_0. A p-value of 1.00 means 0 % chance of a difference, while a p-value of 0.95 means a chance of a difference close to 0 %. A p-value of > 0.95 literally means that we have > 95 per cent chance of finding a result less close to H_0, which means a chance of < (1-0.95), i.e., < 0.05 of finding a result this close or closer. Using the traditional 5 per cent decision level, this would mean, that we have a strong argument that such data are closer to H_0 than compatible with random sampling.
3. An observed standard deviation (sd) < 50% the sd expected from prior population data. From population data we can be pretty sure about sds to be expected. E.g., the sds of blood pressures are close to 10% of their means, meaning that for a mean systolic blood pressures of 150 mm H_g the expected sd is close to 15 mm H_g, for a mean diastolic blood pressure of 100 mm H_g the expected sd is close to 10 mm H_g. If such sds can be assumed to follow a normal distribution, we will have < 5% chance of finding sds < 7.5 and < 5 mm H_g respectively.
4. An observed standard deviation (sd) > 150% the sd expected from prior population data. With sds close to 10% of their means, we, likewise, will have < 5% chance of finding sds > 150% the size of the sds expected from population data.

2. Studies with multiple endpoints

A simple method to check the accuracy of multiple endpoints is to examine the distribution of the final digits of the results, using the chi-square goodness of fit test. In a clinical trial of cholesterol lowering treatment the results were presented mainly in the form of relative risks (RR = risk of mortality during treatment / risk of mortality during control). In total 96 RRs were presented with many of them showing a 9 or 1 as final digit. E.g., RRs of 0.99, 0.89, 1.01, and 1.11 etc were often reported. The accuracy of the multiple endpoints is checked according to Table 2.

Table 2. Multiple risk ratios as reported in a "statin" paper

Final digit of RR	observed frequency	expected frequency	$\sum \frac{(\text{observed-expected})^2}{\text{expected}}$
0	24	9.6	21.6
1	39	9.6	90.0
2	3	9.6	4.5
3	0	9.6	9.6
4	0	9.6	9.6
5	0	9.6	9.6
6	0	9.6	9.6
7	1	9.6	7.7
8	2	9.6	6.0
9	27	9.6	31.5
Total	96	96.0	199.7

If there were no tendencies to record only whole RRs, we would expect equal numbers of 0s, 1s, 2s,….9s for the final digit, that is 9.6 of each. The agreement between the observed and expected digits is, then, tested according to

$$\text{Chi-square} = \sum \frac{(\text{observed-expected})^2}{\text{expected}} = 199.7 \text{ for 10-1 degrees of freedom}$$

(there are 10 different frequencies). For the given degrees of freedom a chi-square value > 27.88 means that the null-hypothesis of no-difference-between-observed-and-expected can rejected at a p-value < 0.001. The distribution of the final digits of the RRs in this study does not follow a random pattern. The presence of unrandomness in these results can be assumed with confidence, and jeopardizes the validity of this study.

4. DISCUSSION

This chapter gives some simple statistical methods to assess trial data for their compliance with the random property. We should add that distribution-free statistical tests that are less dependent on random distributions, are available, but, in practice, they are used far less frequently than normal tests. Also, with slight departures from the normal distribution, normal tests are used even so. The same applies to the analysis of unrandomized studies: for their statistical analysis the same statistical tests are applied as those applied for randomized studies, although this, at the same time, is one of the main limitations of this kind of research. The issue of the current chapter is not the statistical analysis of unrandom data but rather the detection of it.

Regarding the studies with multiple endpoints, the problem is often a dependency of the endpoints in which case the presented method for the assessment of unrandomness is not adequate. E.g. endpoints like deaths, metastases, local relapses, etc in an oncology trial cannot be considered entirely independent. However, if the initial digits of the results are equally distributed, like, e.g., RRs 1.1 / 2.1 / 3.1 / 4.1 / 5.1, then little dependency is to be expected, and the presented method can be properly performed.

Important causes for unrandomness in clinical trials include extreme exclusion criteria[4] and inappropriate data cleaning.[1] The first cause can be illustrated by the study of Kaariainen et al [7] comparing the effect of strict and loose inclusion criteria on treatment results in 397 patients hospitalized for gastric ulcers. While under the loose inclusion criteria virtually none of patients had to be excluded, 71% of them had to be excluded under the strict inclusion criteria. Major complications of treatment occurred in 71 out of the 397 patients with the loose , in only two out of 115 patients with the strict inclusion criteria. These two major complications can hardly be considered representative results from a sample drawn at random from the target population. The second cause can be illustrated by Mendel's pea data. In 1860 Gregor Mendel performed randomized trials "avant la lettre" by using aselective samples of peas with different phenotypes. When we recently looked into Mendel's pea data, and performed a chi-square test, we had to conclude that the chi-square value was too small not to reject the null hypothesis ($P>0.99$).[3] This means that Mendel's reported data were so close to what he expected that we could only conclude that he somewhat misrepresented the data.

The current chapter is an effort to provide the scientific community with some simple methods to assess randomness of experimental data. These methods are routinely used in accountancy statistics for assessing the possibility of financial fraud, but they cannot be found in most textbooks of medical statistics.

Evidence-based medicine is under pressure due to the conflicting results of recent trials producing different answers to similar questions.[8,9] Many causes are mentioned. As long as the possibility of unrandom data has not been addressed, this very possibility cannot be excluded as potential cause for the obvious lack of homogeneity in current research.

5. CONCLUSIONS

Well-balanced randomly sampled representative experimental data comply with the random property meaning that they follow statistical frequency distributions. We continually make use of these frequency distributions to analyze the data, but virtually never assess how close to the expected frequency distributions the data actually are. Unrandom data may be one of the reasons for the lack of homogeneity in the results from current research. The objective of this chapter was to propose some methods for routinely assessing clinical data for their compliance with the random property.

If the individual data from the trial are available, the chi-square goodness of fit and the Kolmogorov-Smirnov goodness of fit tests can be applied (both tests yield similar results and can be applied with any kind of data including continuous data, proportions, or frequencies), for survival data logarithmic transformation can be applied. If the individual data from the trial are not available, the following criteria may be used: observed p-values between 0.0001 and 0.95, observed standard deviations (sds) between 50% and 150% of the sd expected from population data. With multiple endpoints, the distribution of the final digits of the results may be examined using a chi-square goodness of fit test. In the current chapter some simple statistical tests and criteria are given to assess randomized clinical trials for their compliance with the random property.

6. REFERENCES

1. Cleophas TJ. Research data closer to expectation than compatible with random sampling. Stat Med 2004; 23: 1015-17.
2. Cleophas TJ. Clinical trials and p-values, beware of the extremes. Clin Chem Lab Med 2004; 42: 300-4.
3. Cleophas TJ, Cleophas GM. Sponsored research and continuing medical education. J Am Med Assoc 2001; 286: 302-4.
4. Furberg C. To whom do the research findings apply? Heart 2002; 87: 570-574.
5. Levin RI, Rubin DS. Statistics for management. 7th Edition. Edited by Prentice-Hall International, New Jersey, 1998.
6. Hung HM, O'Neill RT, Bauer P, Köhne K. The behavior of the p-value when the alternative hypothesis is true. Biometrics 1997; 53: 11-22.
7. Kaarainen I, Sipponen P, Siurala M. What fraction of hospital ulcer patients is eligible for prospective drug trials? Scand J Gastroenterol 1991; 26: 73-6.
8. Julius S. The ALLHAT study: if you believe in evidence-based medicine, stick to it. J Hypertens 2003; 21: 453-4.
9. Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. Int J Clin Pharmacol Ther 2003; 41: 51-6.

CHAPTER 30

CLINICAL TRIALS DO NOT USE RANDOM SAMPLES ANYMORE

1. INTRODUCTION

Current clinical trials do not use random samples anymore. Instead, they use convenience samples from selected hospitals, including only patients with strict characteristics, like cut-off laboratory values. This practice, although it improves the precision of the treatment comparison, raises the risk of non-normal data. This is a problem since the assumption of normality underlies many statistical tests. If this assumption is not satisfied, the logic underlying the distributions of the test statistics used to estimate whether the observed effects are due to chance rather than treatment effect, fails, and, consequently, the resulting p-values are meaningless. Evidence-based medicine is under pressure due to the heterogeneity of current trials.[1-4] The possibility of non-normal data cannot be excluded as a contributing cause for this. The current paper reviews and describes for a non-mathematical readership methods to assess data for compliance with normality, and summarizes solutions for the analysis of non-normal data.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 367–374.

2. NON-NORMAL SAMPLING DISTRIBUTIONS, GIVING RISE TO NON-NORMAL DATA

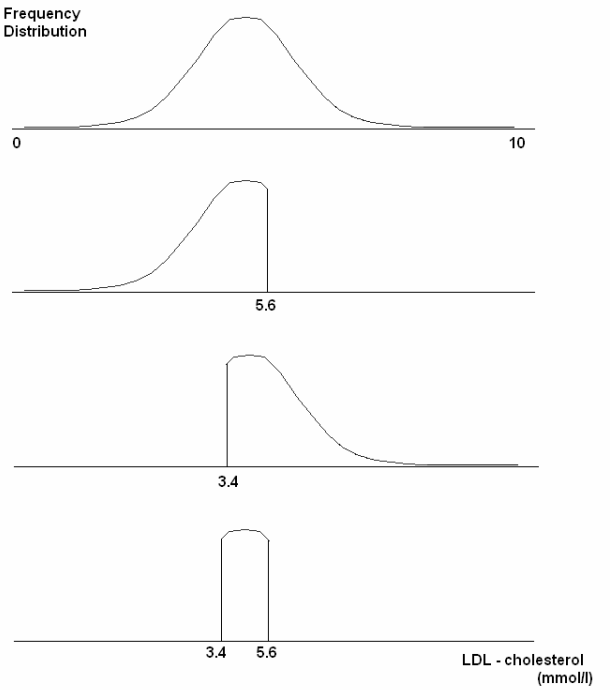


Figure 1. Sampling distributions of patients with heterozygous hypercholesterolemia with on the x-axis individual results and on the y=axis "how often": (1) all of the patients that genetically qualify are included; (2) patients with an LDL-cholesterol ≤ 5.6 included; (3) patients with LDL-cholesterol >3.4 mmol/l included; (4) only patients with LDL-cholesterol between 3.4 and 5.7 mmol/l included.

Figure 1 gives an example of sampling distributions of patients with heterozygous hypercholesterolemia. If all of the patients who genetically qualify are included, we will obtain an entirely normal frequency distribution of the individual LDL-cholesterol values. If, however, only patients with an LDL-cholesterol ≤ 5.7 or >3.4 mmol/l or between 3.4 and 5.7 mmol/l are included[5], we will obtain non-normal distributions as shown in the Figure 1. Figure 2 gives an example of patients with constitutional constipation before and after treatment with a laxative.[6] Only patients with < 3 stools per week were included. The figure 2 shows how a skewed sampling distribution may give rise to a non-normal trial result. This, obviously, happens not only to control groups and comparisons versus baseline,

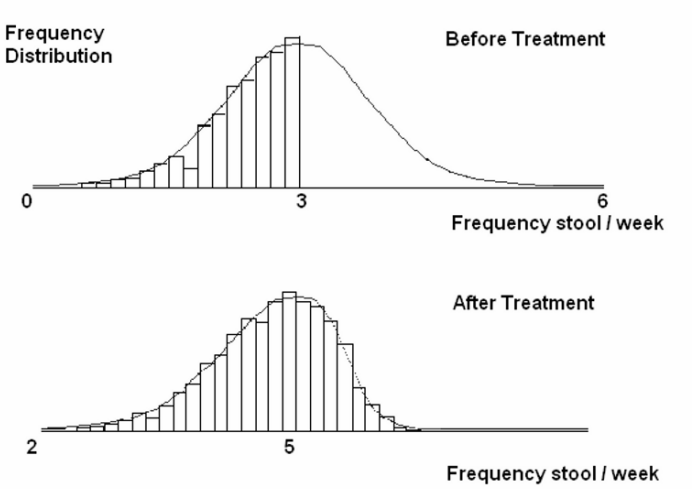


Figure 2. Frequency distributions of patients with constitutional constipation before and after treatment with a laxative. Only patients with < 3 stools per week were included.

but also to active treatment groups, probably, due to the positive correlation between repeated observations in one subject.

3. TESTING THE ASSUMPTION OF NORMALITY

In trials with a single endpoint the chi-square goodness of fit test as discussed in Chapter 29 can be applied. In case of multiple endpoints a simple method to check the normality of multiple endpoints is to examine the distribution of the final digits of the results, also using the chi-square goodness of fit test. The example was modified from Kirkwood and Stern with permission from the editor.[8] In a clinical trial of cholesterol lowering treatment the results were presented mainly in the form of relative risks (RR = risk of mortality during treatment / risk of mortality during control). In total 96 RRs were presented with many of them showing a 9 or 1 as final digit. E.g., RRs of 0.99, 0.89, 1.01, and 1.11 etc were often reported. The accuracy of multiple endpoints is checked as follows:

Final digit of RR	observed frequency	expected frequency	$\sum$ (observed-expected)^2 / expected
0	24	9.6	21.6
1	39	9.6	90.0
2	3	9.6	4.5
3	0	9.6	9.6
4	0	9.6	9.6

5	0	9.6	9.6
6	0	9.6	9.6
7	1	9.6	7.7
8	2	9.6	6.0
9	27	9.6	31.5
Total	96	96.0	199.7

If there were no tendencies to record only whole RRs, we would expect equal numbers of 0s,1s,2s,….9s for the final digit, that is 9.6 of each. The agreement between the observed and expected digits is, then, tested according to chi-square = $\sum$ (observed-expected)^2 / expected = 199.7 for 10-1 degrees of freedom (there are 10 different frequencies). For the given degrees of freedom a chi-square value > 27.88 means that the null-hypothesis of no-difference-between-observed-and-expected can rejected at a p-value < 0.001. The distribution of the final digits of the RRs in this study does not follow a normal pattern. The presence of unrandomness in these results can be assumed, and jeopardizes the validity of this study.

4. WHAT TO DO IN CASE OF NON-NORMALITY

If the data are not normally distributed, they may be given ranknumbers and these ranknumbers often tend to look like normal distributions. For their analysis ranksum tests have been developed, including the Mann-Whitney's and the Wilcoxon's tests. In order to assess whether the data are suitable for rank-testing a special goodness of fit test is available, the Kolmogorov-Smirnov test. as discussed in Chapter 29.
Also confidence intervals based on percentiles can be used if rank-testing is not warranted. Confidence intervals can be derived from the data without prior assumption about the type of distribution. The simplest way to do so is to take the range within which 95% of all possible outcomes lie. Medians rather than means should be used for calculation, because these data are skewed. The underneath method is correct for finding the 95% confidence interval of a difference between medians.

group 1	group 2
3.99	3.18
3.79	2.84
3.60	2.90
3.73	3.27
3.21	3.85
3.60	3.52
4.08	3.23
3.61	2.76
3.81	3.60
median	

3.73 3.23
difference in medians
0.50

All possible differences between the two groups (9 x 9 = 81) lie between –0.64 and 1.24. After exclusion of 2.5% of the lowest and 2.5% of the highest differences, we will obtain a range containing 95% of the differences. This 95% confidence interval is between –0.25 and 1.15. This interval includes the value 0, which means that the difference between the two groups is not significantly different from 0 ($p > 0.05$).

A largely similar but more sophisticated method to obtain confidence intervals from your data is the bootstrap method.[9] Bootstrapping, otherwise called jack-knifing, is a data based simulation process for statistical inference. The basic idea is sampling with replacement in order to produce random samples from the original data. The procedure is illustrated underneath for two bootstrap samples. In the first bootstrap sample observation 1 was picked up twice, while observations 2 and 4 were not picked. We repeat this procedure a large number of times, and record the difference in medians of each bootstrap sample. To derive confidence intervals, at least $n_{group\ 1} \times n_{group\ 2} = 9 \times 9 = 81$ bootstraps are required. To calculate the confidence intervals the percentile method like described above can be used. In this example a 95 % confidence interval between –0.12 and 1.09 was obtained, again not significantly different from 0.

Original data		bootstrap 1		bootstrap 2	
group1	group 2	group1	group 2	group 1	group 2
1. 3.99	10. 3.18	1. 3.99	10. 3.18	1. 3.99	10. 3.18
2. 3.79	11. 2.84	1. 3.99	10. 3.18	2. 3.79	11. 2.84
3. 3.60	12. 2.90	3. 3.60	12. 2.90	2. 3.79	12. 2.90
4. 3.73	13. 3.27	5. 3.21	14. 3.85	2. 3.79	12. 2.90
5. 3.21	14. 3.85	6. 3.60	15. 3.52	4. 3.73	14. 3.85
6. 3.60	15. 3.52	8. 3.61	15. 3.52	5. 3.21	15. 3.52
7. 4.08	16. 3.23	8. 3.61	15. 3.52	7. 4.08	16. 3.23
8. 3.61	17. 2.76	9. 3.81	16. 3.23	7. 4.08	18. 3.60
9. 3.81	18. 3.60	9. 3.81	17. 2.76	8. 3.61	18. 3.60
median group1 = 3.73		median group 1 = 3.61		median group 1 = 3.79	
median group2 = 3.23		median group 2 = 3.23		median group 2 = 3.23	
difference medians 0.50		difference medians = 0.38		difference medians = 0.55	

5. DISCUSSION

Current clinical trials do not, usually, use random samples, but rather convenience samples from selected hospitals, including only patients with strict characteristics, like cut-off laboratory values. This practice raises the risk of non-normal data. The assumption of normality underlies many statistical tests, including, among other tests, normal-, t-, chi-square- tests, analysis of variance, and regression analyses. With slight departures from the normal distribution, these tests can be used even

so. They should not be used, however, if the chi-square goodness of fit is significant. Rank-testing is, then, an adequate alternative. But, sometimes, distributions do not allow for this approach either. This can be checked by the Kolmogorov-Smirnov test. If the latter test is also positive, rank-testing is not warranted, and confidence intervals can be derived from the data without prior assumption about the type of frequency distribution. This can be done by calculating the range within which 95% of all possible outcomes. We should add that medians, instead of means, are recommended for calculation, because the data are skewed. Another popular method for this purpose is bootstrapping, which resamples at random from the study's own data. The basic idea is simple: if we take repeated samples from the data themselves, we will mimick the way the data were sampled from the population in the way it should, namely at random. Although the theoretical properties of bootstrapping have not been well-understood, practical performance of this resampling method has been demonstrated in a number of simulation studies.[10]

The current paper is just an introduction, and many aspects are not covered. SPSS[11] uses, instead of the chi-square goodness of fit test, the Shapiro-Wilk[12] test, which is mathematically more complicated, but performs otherwise largely similarly. We should add that, in clinical research, some data are, traditionally, non-normal, but have been recognized to follow a normal distribution after transformation. This is true for risk ratios, odds ratios, and hazard ratios that are analyzed after logarithmic transformations by simple normality tests. The final results can, then, be retrieved by taking the antilog term of the obtained result.

Well-balanced randomly sampled representative experimental data often comply with the so-called random property meaning that they follow normal or close to normal frequency distributions. We continually make use of these frequency distributions to analyze the data, but virtually never assess how close to the expected frequency distributions the data actually are. Unrandom data due to convenience sampling and extreme inclusion criteria may be one of the reasons for the lack of homogeneity in current research. We strongly believe that normality statistics, although the mainstay of statistical analysis for centuries, will rapidly be replaced with non-normal testing as the awareness of non-normal sampling distributions grows. Moreover, confidence intervals from data without prior assessment of frequency distributions are, currently, more easy to obtain than in the past, because computers can produce hundreds of random numbers from any set of experimental data within seconds. E.g., with the function RANDBETWEEN, after giving the data ranknumbers, the EXCEL program[13] can produce many random samples from any given population as well as their characteristics like medians, ranges, percentiles. We hope that this paper will strengthen the awareness of non-normal sampling distributions, and affect the design and analysis of future clinical trials.

6. CONCLUSIONS

Current clinical trials do not use random samples, but, instead, convenience samples. This raises the risk of non-normal data. This chapter reviews and describes for a non-mathematical readership common methods for testing the normal property, as well as methods for analyzing the data in case of non-normality.

With slight departures from the normal distribution, normality tests can be used even so. They include, among other tests, the normal-, t-, chi-square- tests, analysis of variance, and regression analyses. They should not be used, if the chi-square goodness of fit is significant. Rank-testing is, then, an alternative, but, sometimes, distributions do not allow for this approach either. This can be checked by the Kolmogorov-Smirnov test. If the latter test is also positive, rank-testing is not warranted, and confidence intervals can be derived from the data without prior assumption about the type of frequency distribution. This can be done by calculating the range within which 95% of all possible outcomes lie. Another popular method for this purpose is bootstrapping, which resamples at random from the study's own data.

This chapter reviews methods to assess data for compliance with normality, and summarizes solutions for the analysis of non-normal data. We, strongly, believe that normality statistics, although the mainstay of statistical analysis for centuries, will rapidly be replaced with non-normal testing as the awareness of non-normal sampling distributions grows, and hope that the paper will strengthen this awareness, and affect the design and analysis of future clinical trials.

7. REFERENCES

1. Cleophas TJ. Research data closer to expectation than compatible with random sampling. Stat Med 2004; 23: 1015-7.
2. Furberg C. To whom do the research findings apply? Heart 2002; 87: 570-4.
3. Kaaraininen I, Sipponen P, Siurala M. What fraction of hospital ulcer patients is eligible for prospective drug trials? Scand J Gastroenterol 1991; 26: 73-6.
4. Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. Int J Clin Pharmacol Ther 2003; 41: 51-6.
5. Anonymous. KOWA Study Protocol. Pharmanet, Pharm@net
6. Cleophas TJ, Zwinderman AH, Cleophas TF. Example of crossover trial comparing efficacy of a new laxative versus bisacodyl. Statistics applied to clinical trials. Springer, New York 3rd edition, 2006, pp 126-9.
7. Levin RI, Rubin DS. Statistics for management. 7th Edition. Edited by Prentice -Hall International, New Jersey, 1998.
8. Kirkwood BR, Sterne JA. Medical statistics. 2ndEdition. Blackwell Science, Oxford, UK, 2003.
9. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which , what? A practical guide. Stat Med 2000; 19: 1141-64.
10. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall, 1993, New York, USA.
11. SPSS Statistical Software 2000 Chicago, IL, USA.

12. Royston P. A toolkit for testing for non-normality in complete and censored samples. The Statistician 1993; 42: 37-43.
13. Excel. Microsoft Office Online: EXCEL 2003 Home Page

CHAPTER 31

CLINICAL DATA WHERE VARIABILITY IS MORE IMPORTANT THAN AVERAGES

1. INTRODUCTION

In clinical studies, efficacies of new treatments are usually assessed by comparing averages of new treatment results versus control or placebo. However, averages do not tell the whole story, and the spread of the data may be more relevant. E.g., when we assess how a drug reaches various organs, variability of drug concentrations is important, as in some cases too little and in other cases dangerously high levels get through. Also, for the assessment of the pharmacological response to a drug, variabilities may be important. E.g., the effects on circadian glucose levels in patients with diabetes mellitus of a slow-release-insulin and acute-release-insulin formula are different. The latter formula is likely to produce more low and high glucose levels than the former formula. Spread or variability of the data is a determinant of diabetic control, and predictor of hypoglycaemic / hyperglycemic events. As an example, in a parallel-group study of n = 200 the former and latter formulas produced mean glucoses of 7.9 and 7.1 mmol/l, while standard deviations were 4.2 and 8.4 mmol/l respectively. This suggests that, although the slow-release formula did not produce a better mean glucose level, it did produce a smaller spread in the data. How do we test these kinds of data. Clinical investigators, although they are generally familiar with testing differences between averages, have difficulties testing differences between variabilities. The current chapter gives examples of situations where variability is more relevant than averages. It also gives simple statistical methods for testing such data. Statistical tests comparing mean values instead of variabilities are relatively simple and are one method everyone seems to learn. It is a service to the readership of this book to put more emphasis on variability.

2. EXAMPLES

TESTING DRUGS WITH SMALL THERAPEUTIC INDICES

Aminoglycosides like gentamicin and tobramicin are highly efficaceous against gram-negative bacilli, even pseudomonas. However, their therapeutic indices are small, and, particularly, irreversible nephrotoxicity requires careful monitoring of high plasma levels, while low levels lack therapeutic efficacy.[1] For efficacy/safety

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 375–384.

assessments of such compounds, in addition to monitoring too high and too low averages, monitoring variability of plasma levels is relevant.

TESTING VARIABILITY IN DRUG RESPONSE

In patients with hypertension the effects on circadian blood pressure levels of blood pressure lowering agents from different classes are different. For example, unlike beta-blockers, calcium channel blockers and angiotensin converting enzyme inhibitors amplified amplitudes of circadian blood pressure rhythms.[2] Spread or variability of the data is a determinant of hypertensive control, and predictor of cardiovascular risks.[3] Particularly, for the assessment of ambulatory blood pressure measurements variability in the data is important.

ASSESSING PILL DIAMETERS OR PILL WEIGHTS

A pill producing device is approved only if it will produce pill diameters with a standard deviation (SD) not larger than, e.g., 7 mm. Rather than the average diameter, the variability of the diameters is required for testing the appropriateness of this device.

COMPARING DIFFERENT PATIENT GROUPS FOR VARIABILITY IN PATIENT CHARACTERISTICS

Anxious people may not only show a lower average of performance, but also a higher variability in performance relative to their non-anxious counterparts. Variability assessment is required to allow for predictions on performances.

ASSESSING THE VARIABILITY IN DURATION OF CLINICAL TREATMENTS

For hospital managers the variability in stay-days in hospital is more relevant than the mean stay-days, because greater variability is accompanied with a more demanding type of care.

FINDING THE BEST METHOD FOR PATIENT ASSESSMENTS

A clinician needs to know whether variability in rectal temperature is larger than variability in oral temperature in order to choose the method with the smaller variability.

Various fields of research, particularly in clinical pharmacology, make use of test procedures that, implicitly, address the variability in the data. For example, bioavailability studies consider variability through individual and population bioequivalence instead of just averages.[4,5] For the assessment of diagnostic estimators, repeatability tests and receiver operating (ROC) curves are applied.[6] Mann-Whitney tests for repeated measures consider whether treatment A is better than B.[7] However, none of such studies are especially designed to test variability. The current chapter reviews statistical methods especially designed for such purpose.

3. AN INDEX FOR VARIABILITY IN THE DATA

Repeated observations exhibit a central tendency, the mean, but, in addition, exhibit spread or dispersion, the tendency to depart from the mean. If measurement of central tendency is thought of as good bets, then measures of spread represent the

poorness of central tendency, otherwise called deviation or error. The larger such deviations are, the more do cases differ from each other and the more spread does the distribution show. For the assessment of spread in the data we need an index to reflect variability. First of all, why not simply express variability as the departures of the individual data from the mean value. This, however, will not work, because the data will produce both negative and positive departures from the mean, and the overall variability will approach zero. A device to get around this difficulty is to take the add-up sum of the squares of deviations from the mean, and divide by n-1 (n= sample size):

$$\frac{(\text{datum 1} - \text{mean})^2 + (\text{datum 2} - \text{mean})^2 + (\text{datum 3} - \text{mean})^2 + \ldots}{\text{n -1}}$$

This formula presents the variance of n observations, and is widely used for the assessment of variability in a sample of data. The use of "n-1" instead of "n" for denominator is related to the so-called degrees of freedom. Variances can be applied to assess data like those given in the above examples. The following tests are adequate for such purpose: the chi-square test for a single sample, the F-test for two samples, and the Bartlett's or Levene's tests for three or more samples. Additional methods for analyzing variability include (1) comparisons of confidence intervals, and (2) testing confidence intervals against prior defined intervals of therapeutic tolerance or equivalence. We should add that the variance is only one way to measure variability. Median absolute deviation (MAD) is another method not uncommonly used for pharmaceutical applications. It is found by taking the absolute difference of each datum from the sample median, and, then, taking the median of the total number of values. MADs will not be further discussed in this chapter.

4. HOW TO ANALYZE VARIABILITY, ONE SAMPLE

1. χ^2 test

For testing whether the standard deviation (or variance) of a sample is significantly different from the standard deviation (or variance) to be expected the chi-square test is adequate. The chi-square test is closely related to the normal test or the t-test. The main difference is the use of squared values in the former. The underneath formula is used to calculate the chi-square value

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \text{for n-1 degrees of freedom}$$

(n = sample size, s = standard deviation, s^2 = variance sample, σ = expected standard deviation, σ^2 = expected variance).

For example, in an ongoing quality control produced tablets are monitored for consistency in size. Samples of 50 tablets are only approved if the sample size

standard deviation value is less than 0.7 mm. A 50 tablet sample has a standard deviation of 0.9 mm.

$$\chi^2 = (50\text{-}1)\ 0.9^2 / 0.7^2 = 81$$

The chi-square table shows that, for 50-1 = 49 degrees of freedom, we find a p-value <0.01 (one-sided test). This sample's standard deviation is significantly larger than that required. This means that this sample cannot be approved.

2. Confidence interval

Instead of, or in addition to, the above chi-square test a confidence interval can be calculated. It can be more relevant than, simply, the above test, and it is considered good statistical practice to provide a confidence interval to accompany any estimates. The underneath formulas are used for calculation.

$$s\sqrt{\frac{(n-1)}{b}} \text{ and } s\sqrt{\frac{(n-1)}{a}}$$

n = sample size, s = standard deviation
b = cut-off value of left tail of χ^2 – distribution for given α and degrees of freedom
a = cut-off value of right tail of χ^2 – distribution for given α and given degrees of freedom
α = type I error

We use the above example, with a standard deviation (s) of 7 mm and observed s of 9 mm, to calculate 90% confidence interval (α = 10%). As the sample size = n = 50, the degrees of freedom is n-1 = 49. The cut-off values, b and a, can be found in the left and right tail χ^2 tables, available in statistics textbooks, statistical software, and literature.[8]

$$s\sqrt{\frac{(n-1)}{b}} = 9\times\sqrt{\frac{49}{63.17}} = 9\times 0.88\text{mm} \quad \text{and}$$
$$s\sqrt{\frac{(n-1)}{b}} = 9\times\sqrt{\frac{49}{37.69}} = 9\times 1.14\text{mm}$$

The 90% confidence interval is, thus, between 8.1 and 10.1 mm, and it does not cross the required standard deviation of 7 mm. The device is not approved.

3. Equivalence test

A limitation of the above methods is that a statistical difference is demonstrated using standard deviations. To clinical pharmacologists a more appealing approach might be an equivalence test which uses prior defined boundaries of equivalence,

and, subsequently, tests whether the 90 or 95 % confidence intervals of a sample are within these boundaries. If entirely within, we accept equivalence, if partly within we are unsure, if entirely without we conclude lack of equivalence. Furthermore, what is nice about equivalence intervals, is, that both mean and variability information are incorporated. Basic references are the guidelines given by Schuirmann and Hahn.[9,10] As an example, the boundaries for demonstrating equivalence of the diameters of a pill could be set between 9.0 and 11.0 mm. A pill producing device produces a sample with a mean diameter of 9.8 mm and 90% confidence intervals of ± 0.7 mm. This would mean that the confidence intervals are between 9.1 and 10.5 mm, and that they are, thus, entirely within the set boundary of equivalence. We can state that we are 90% confident that at least 90% of the values lie between 9.1 and 10.5 mm (type I error 10%). According to this analysis, the pill producing device can be approved.

5. HOW TO ANALYZE VARIABILITY, TWO SAMPLES

1. F test

F tests can be applied to test if variability of two samples is significantly different. The division sum of the samples' variances (larger variance / smaller variance) is used for the analysis. For example, two formulas of gentamicin produce the following standard deviations of plasma concentrations:

	Patients (n)	standard deviation (s) (µg/l)
formula-A	10	3.0
formula-B	15	2.0

$F = s_{Formula\text{-}A}^2 / s_{Formula\text{-}B}^2 = 3.0^2 / 2.0^2 = 9 / 4 = 2.25$
with degrees of freedom (dfs) for formula-A 10-1=9 and for formula-B 15-1 = 14.

The F-table shows that an F-value of at least 3.01 is required not to reject the null - hypothesis. Our F-value is 2.25 and, so, the p-value is >0.05. No significant difference between the two formulas can be demonstrated. This F-test is available in Excel.

2. Confidence interval

Also for two samples the calculation of confidence intervals is possible. It will help to assess to what extent the two formulations actually have similar variances or whether the confidence interval is wide and, thus, the relationship of the two variances is really not known. The formulas for calculation are given.

(1 / cut-off F-value) x calculated F-value and
(cut-off F-value) x calculated F-value

Cut-off F-value = F-value of F-table for given α and degrees of freedom
α = type I error

We calculate the 90% confidence interval from the above two sample example.

(1 / cut-off F-value) x calculated F-value = 1/ 3.01 x 2.25 = 0.75 and
(cut-off F-value) x calculated F-value = 3.01 x 2.25 = 6.75

The 90% confidence interval for this ratio of variances is between 0.75 and 6.75. This interval crosses the cut-off F-value of 3.01. So, the result is not significantly different from 3.01. We conclude that no significant difference between the two formulations is demonstrated.

3. Equivalence test

An equivalence test for two variances works largely similar to a therapeutic tolerance interval test for a single variance. We need to define a prior boundary of equivalence, and then, test whether our confidence interval is entirely within. A problem with ratios of variances is that they often have very large confidence intervals. Ratios of variances are, therefore, not very sensitive to test equivalence. Instead, we can define a prior overall boundary of equivalence and, then, test whether either of the two variances is within. E.g., in the above two variances example the boundary of equivalence of plasma concentration of gentamicin for 90 % confidence intervals had been set between 3.0 and 7.0 µg/l. The mean plasma concentrations were 4.0 for formula-A and 4.5 µg / l for formula-B.

Patients (n)		standard (s) deviation (µg / l)	mean (µg / l)	standard error	90% confidence interval
formula-A	10	3.0	4.0	0.9	2.5 to 5.5
formula-B	15	2.0	4.5	0.6	3.5 to 5.5

As the 90 % confidence interval for formula-A is not entirely within the set boundary, the criterion of equivalence is not entirely met. Based on this analysis, equivalence of the two formulas cannot be confirmed.

6. HOW TO ANALYZE VARIABILITY, THREE OR MORE SAMPLES

1. Bartlett's test

Bartlett's test can be applied for comparing variances of three samples

$$\chi^2 = (n_1 + n_2 + n_3 - 3)\ln s^2 - [(n_1 - 1)\ln s_1^2 + (n_2 - 1)\ln s_2^2 + (n_3 - 1)\ln s_3^2]$$

where n_1 = size sample 1
s_1^2 = variance sample 1

$$s^2 = \text{pooled variance} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n_1 + n_2 + n_3 - 3} =$$

ln = natural logarithm

As an example, blood glucose variabilities are assessed in a parallel-group study of three insulin treatment regimens. For that purpose three different groups of patients are treated with different insulin regimens. Variabilities of blood glucose levels are estimated by group-variances:

	group size (n)	variance [(mmol/l)^2]
Group1	100	8.0
Group 2	100	14.0
Group 3	100	18.0

Pooled variance = $\frac{99 \times 8.0 + 99 \times 14.0 + 99 \times 18.0}{297} = 13.333$

χ^2 = 297 x ln 13.333 – 99 x ln 8.0 - 99 x ln 14.0 – 99 x ln 18.0 =
297 x 2.58776 – 99 x 2.079 – 99 x 2.639 – 99 x 2.890 =
768.58 – 753.19 =
15.37

We have three separate groups, and, so, 3-1 = 2 degrees of freedom. The chi-square table shows that a significant difference between the three variances is demonstrated at $p<0.001$. If the three groups are representative comparable samples, we may conclude that these three insulin regimens do not produce the same spread of glucose levels. In this study of parallel groups, variability in the data is assessed by comparison of between-subject variability. Other studies assess variability in the data by repeated measures within one subject.

2. Levene's test

An alternative to the Bartlett's test is the Levene's test. The Levene's test is less sensitive than the Bartlett's test to departures from normality. If there is a strong evidence that the data do in fact come from a normal, or nearly normal, distribution, then Bartlett's test has a better performance. Both tests can be used for comparison of more than two variances. However, we should add that assessing significance of differences between more than two variances is, generally, not so relevant in clinical comparisons. In practice, clinical investigators are mostly interested in differences between two samples / groups rather than multiple samples / groups.

7. DISCUSSION

For all tests discussed above we need to emphasize that the data come from a normal distribution. The tests can be quite misleading if applied to non-normal data. It would be good practice to look at the distribution of the data first, for example by drawing histograms or box plots, and to transform if needed. Also, non-parametric tests are available for the analysis of variances of non-normal data, for example the Kendall's test for the variance of ranks.[11-13]

In the current paper eight statistical methods are described for comparing variances of studies where the emphasis is on variability in the data. Clinical examples are given. The assessment of variability is, particularly, important in studies of medicines with a small therapeutic index. Table 1 gives an overview of such medicines, commonly used in practice. Their therapeutic ranges have been defined, and it is a prerequisite of many of them that peak and trough concentrations are carefully monitored in order to reduce toxicities and improve therapeutic efficacies. The development of such therapeutic ranges can benefit from variance-testing. For other medicines therapeutic indices may not be small, while plasma concentrations are not readily available. Instead of dose-concentration relationships, dose-response relationships are, then, studied in order to determine the best therapeutic regimens. This approach uses dose-response curves, and is based on the assumption that the mean response of many tests can be used for making predictions for the entire population. However, dose-response relationships may differ between individuals, and may depend on determinants like body mass, kidney function, underlying diseases, and other factors hard to control. Moreover, for the treatment of diseases like diabetes mellitus, hypercholesterolemia, hypertension etc, we are often more interested in the range of responses than we are in the mean response. Also for the study of such data variance-testing would, therefore, be in place.

Table 1. Drugs with small therapeutic indices

1. Antibacterial agents
gentamicin, vancomicin, tobramicin
2. Drugs for seizure disorders
carbamazepine, phenytoine, phenobarbital, valproate
3. Cardiovascular and pulmonary drugs
digoxin, theophylline, caffeine
4. Antidepressant drugs
amitryptiline, nortriptyline, imipramine,clomipramine,maprotiline
5. Neuroleptic drugs
clozapine

Samples of observations are unpaired, if every patient is tested once, or paired, if every patient is tested repeatedly. In the case of repeated testing special statistical procedures have to be performed to adjust for correlations between paired observations. This is, particularly, required when analyzing averages, but less so when analyzing variances. Correlation levels little influence the comparison of variances, and, so, similar tests for the comparison of variances can be adequately used both for paired and for unpaired variances.

In conclusion, in clinical studies variability of the data may be a determinant more important than just averages. The current chapter provides eight straightforward methods to assess normally distributed data for variances, that can be readily used. The chi-square test for one sample and the F-test for two samples are available in Excel. The Bartlett's and Levene's test can be used for multiple variances, and are not in Excel, but can be found in statistical software programs. For the readers' convenience a reference is given.[14] Also, references are given for methods to analyze variances from non-normal data.[11-13]

8. CONCLUSIONS

Clinical investigators, although they are generally familiar with testing differences between averages, have difficulty testing differences between variabilities. The objective of this chapter was to give examples of situations where variability is more relevant than averages. Also to give simple methods for testing such data.
Examples include: (1) testing drugs with small therapeutic indices, (2) testing variability in drug response, (3) assessing pill diameters or pill weights, (4) comparing patient groups for variability in patient characteristics, (5) assessing the variability in duration of clinical treatments, (6) finding the best method for patient assessments. Various fields of research, particularly in clinical pharmacology, make use of test procedures that, implicitly, address the variability in the data. Tests especially designed for testing variabilities in the data include chi-square tests for one sample, F-tests for two samples, and Bartlett's or Levene's tests for three or more samples. Additional methods include (1) comparisons of confidence intervals, and (2) testing confidence intervals against prior defined intervals of therapeutic tolerance or equivalence. Many of these tests are available in Excel, and other statistical software programs, one of which is given.
We conclude that for the analysis of clinical data the variability of the data is often more important than the averages. Eight simple methods for assessment are described. It is a service to the readership of this book to put more emphasis on variability.

9. REFERENCES

1. Chambers HF, Sande MA. Antimicrobial agents: the aminoglycosides. In Goodman and Gillman's Pharmacological basis of therapeutics, 9th edition, McGraw Hill, New York, USA, 1996; pp 1103-21

2. Cleophas TJ, Van der Meulen J, Zwinderman AH. Nighttime hypotension in hypertensive patients prevented by beta-blockers but not by angiotensin converting enzyme inhibitors or calcium channel blockers. Eur J Intern Med 1998; 9: 251-7.
3. Neutel JM, Smith DH. The circadian pattern of blood pressure: cardiovascular risk and therapeutic opportunities. Curr Opin Nephrol Hypertens 1997; 6: 250-6.
4. Hauck WW, Anderson S. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. J Pharmacokin Biopharmaceuticals 1984; 12: 83-91.
5. Tothfalusi L, Endrenyl L. Evaluation of some properties of individual bioequivalence from replicate-design studies. Int J Clin Pharmacol Ther 2001; 39: 162-6.
6. Almirall J, Bolibar I, Toran P, et al. Contribution of C-reactive protein to the diagnosis and assessment of severity of community-acquired pneumonia. Chest 2004; 125: 1335-42.
7. Petrie A, Sabin C. Medical statistics at a glance. London, UK, Blackwell Science Ltd, 2000.
8. Cleophas TJ. Statistical tables to test data closer to expectation than compatible with random sampling. Clin Res Reg Affairs 2005; 22: 83-92.
9. Schuirmann DJ. A comparison of the two one-sided test procedures and the proper approach for assessing the equivalence of average bioavailability. J Pharmacokin Biopharmaceuticals 1987; 15: 657-80.
10. Hahn GJ, Meeker WQ. Statistical intervals: a guide for practitioners. New York, John Wiley and Sons, Inc, 1991.
11. Kendall MG, Stuart A. Rank correlation methods. 3rd edition, London, UK, Griffin, 1963.
12. Siegel S. Non-parametric methods for behavioural sciences. New York, McGraw Hill, 1956.
13. Tukey JW. Exploratory data analysis. Reading, MA, Addison-Wesley, 1977.
14. Anonymous http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

CHAPTER 32

TESTING REPRODUCIBILITY

1. INTRODUCTION

Poor reproducibility of diagnostic criteria is seldom acknowledged as a cause for low precision in clinical research. Yet, very few clinical reports communicate the levels of reproducibility of the diagnostic criteria they use. For example, of 11 - 13 original research papers published per issue in the 10 last 2004 issues of the journal Circulation, none did, and of 5 - 6 original research papers published per issue in the 10 last 2004 issues of the Journal of the American Association only one out of 12 did. These papers involved quality of life assessments, which are, notoriously, poorly reproducible. Instead, many reports used the averages of multiple measurements in order to improve precision without further comment on reproducibility. For example, means of three blood pressure measurements, means of three cardiac cycles, average results of morphometric cell studies from two examiners, means of 5 random fields for cytogenetic studies were reported. Poor reproducibility of diagnostic criteria is, obviously, a recognized but rarely tested problem in clinical research. Evidence-based medicine is under pressure due to the poor reproducibility of clinical trials.[1,2] As long as the possibility of poorly reproducible diagnostic criteria has not been systematically addressed, this very possibility cannot be excluded as a contributing cause for this. The current paper reviews simple methods for routine assessment of reproducibility of diagnostic criteria / tests. These tests can answer questions like (1) do two techniques used to measure a particular variable, in otherwise identical circumstances, produce the same results, (2) does a single observer obtain the same results when he/she takes repeated measurements in identical circumstances, (3) do two observers using the same method of measurement obtain the same result.

2. TESTING REPRODUCIBILITY OF QUANTITATIVE DATA (CONTINUOUS DATA)

Method 1, duplicate standard deviations (duplicate SDs)

Reproducibility of quantitative data can be assessed by duplicate standard deviations. They make use of the differences between two paired observations. For example, 10 patients are tested twice for their cholesterol-levels (mmol/l), (Figure 1).

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 385–395.

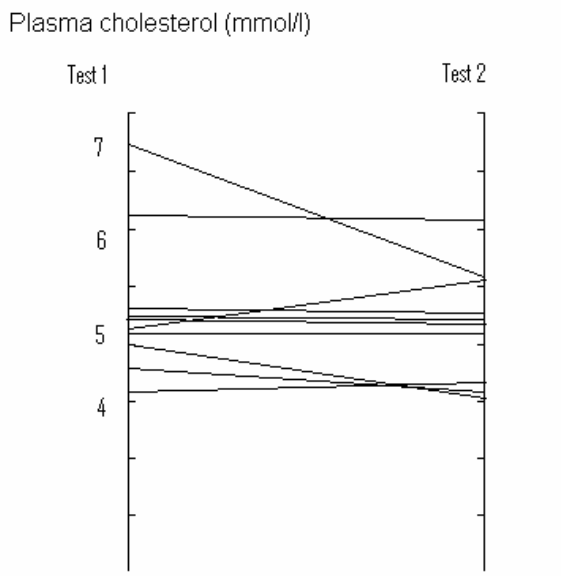


Figure 1 . Ten patients are tested twice for their plasma cholesterol levels

		test-1	test-2	difference (d)	d^2
Patient	1	5.4	5.5	-0.1	0.01
	2	5.5	5.4	0.1	0.01
	3	4.6	4.3	0.3	0.09
	4	5.3	5.3	0.0	0.0
	5	4.4	4.5	-0.1	0.01
	6	5.5	5.4	0.1	0.01
	7	6.6	6.4	0.2	0.02
	8	5.4	5.6	-0.2	0.04
	9	4.7	4.3	0.4	0.16
	10	7.3	5.7	1.6	2.56
mean		5.47	5.24	0.23	0.291
sd		0.892	0.677		

Duplicate standard deviation = $\sqrt{\frac{1}{2}\sum\frac{d^2}{n}} = \sqrt{\frac{1}{2}\times 0.291} = 0.3814$mmol/l

d = differences between first and second measurements
n = sample size

Relative duplicate standard deviation = $\frac{\text{duplicate standard deviation}}{\text{overall mean of data}}$

$= 0.3814 \ / \ [(5.47+5.24) / 2]$

$= 0.0726 = 7.3\ \%$

Method 2, repeatability coefficients

Repeatability coefficients equally make use of the differences between two paired observations.

The repeatability coefficient = 2 standard deviations (sds) of paired differences

$$= 2\sqrt{\sum \frac{(d - \bar{d})^2}{n - 1}} = 1.03$$

d = differences between first and second measurements

$\bar{d}$ = mean difference between first and second measurements

n = sample size

The advantage of the repeatability coefficient is that 95 % limits of agreement can be calculated from it. These are between $\bar{d}$ ± 2 sds = 0.23 ± 1.03 = between – 0.80 and 1.26. Under the assumption of a normal distribution we can expect 95% of the data to lie between these limits (Figure 2).

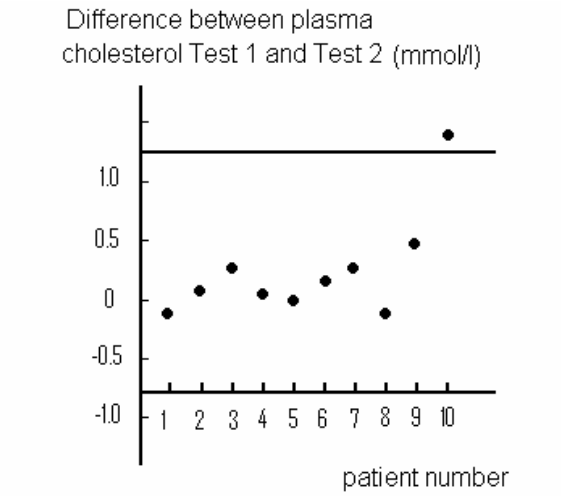


Figure 2. Differences between first and second test for plasma cholesterol in the ten patients from Figure 1. Nine of these ten patients have their differences within the 95 % limits of agreement (two horizontal lines).

Method 3, Intraclass correlation coefficients (ICCS)

Conceptually more complex is the calculation of intraclass correlation coefficients (ICCs) for assessment of between-test agreement. It assesses reproducibility between repeated measures within one subject by comparing the variability between the repeated measures with the total variability in the data.[3] The formula is given by:

$$\text{Intraclass correlation coefficient (ICC)} = \frac{\text{sd}^2 \text{ between subjects}}{\text{sd}^2 \text{ between subjects} + \text{sd}^2 \text{ within subjects}}$$

The ICC ranges from 0 to 1, and it reflects the strength of association between the first and second test. If it equals zero, no reproducibility can be expected. If 1, then reproducibility is 100%. The ICC is otherwise called *proportion of variance* or *correlation ratio*. If you are using SPSS[4] to analyze the data, there is an easy way to calculate the coefficient, which, additionally, provides you with a confidence interval and a p-value. A significant p-value is to be interpreted in terms of a proportion of total variance responsible for between-measurement variation significantly greater than 0. First command: Analyze / Scale / Reliability analysis. The dialog box allows you to include the two variables (results test-1 and results test-2). Next click the statistics box, and select the intraclass correlation coefficient, Model: one-way random, continue. The results for the above example are listed underneath:

Intraclass correlation coefficient = 0.7687
95% confidence intervals between 0.3386 and 0.9361
p-value 0.002
proportion of total variance responsible for between test variability = 77%.

ICCs can also be used for more than two repeated measures.

3. TESTING REPRODUCIBILITY OF QUALITATIVE DATA (PROPORTIONS AND SCORES)

Cohen's kappas

We use the example used by the Colorado Education Geography Center.[5] Suppose two observers assess the same patients for congenital heart disease, using Perloff's classification A to E[6] , and we wish to evaluate the extent to which they agree.

		Observer 1					
		A	B	C	D	E	total
Observer 2	A	**2**	0	2	0	0	4
	B	0	**1**	0	0	0	1
	C	1	0	**1**	0	0	2
	D	0	0	0	**2**	1	3
	E	0	0	0	0	**6**	6
	Total	3	1	3	2	7	16 (N)

We present the results in a two-way contingency table of frequencies. The frequencies with which the observers agree are shown along the diagonal of the table (fat print). Note that all observations would be in the diagonal if they were perfectly matched. Then calculate the q-values, where q = the number of cases expected in the diagonal cells by chance.

$q = n_{row} \times n_{column} / N$
A = 4 x 3 / 16 = 0.75
B = 1 x 1 / 16 = 0.0625
C = 2 x 3 / 16 = 0.375
D = 3 x 2 / 16 = 0.375
E = 6 x 7 / 16 = 2.625
q total = 4.1875 = 4.2

Then calculate kappa:

kappa = (d-q) / (N-q)
d = 12 (the diagonal total of cells = 2+1+1+2+6 =12)
N = total of columns or rows which should be equal
kappa = (12-4.2) / (16-4.2) = 0.66.

The closer the kappa is to 1.0, the better the agreement between the observers:

Poor if	k<0.20
Fair	0.21<k<.040
Moderate	0.41<k<0.60
Substantial	0.61<k<0.80
Good	k>0.80.

4. INCORRECT METHODS TO ASSESS REPRODUCIBILITY

Testing the significance of difference between two or more sets of repeated measures

Instead of the repeatability coefficients or duplicate standard deviations, sometimes the significance of differences between two means or two proportions is used as method to assess reproducibility. For that purpose paired t-tests or McNemar's tests are used. For more than two sets of repeated measures tests like the repeated-measures-analysis-of-variance or Friedman's tests are adequate. As an example, the significance of difference between the above two columns of cholesterol values are calculated as follows (sd = standard deviation, se = standard error):

mean difference ± sd = 0.23 ± 0.5165
mean difference ± se = 0.23 ± 0.1633
t-value = 0.23 / 0.1633 = 1.41
according to the t-table $p > 0.05$

This means that no significant difference between the first and second set of measurements is observed. This can not be taken equal to evidence of reproducibility. With small samples no evidence of a significant difference does not necessarily imply the presence of reproducibility. Yet, a test to preclude a significant difference is relevant within the context of reproducibility statistics, because it establishes the presence of a systematic difference. We are dealing with a biased assessment if we want to test the null-hypothesis of reproducibility.

Calculating the level of correlation between two sets of repeated measures

If you plot the results from the first occasion against those from the second occasion, and calculate a Pearson's regression coefficient, a high level of correlation does not necessarily indicate a great reproducibility. For testing reproducibility we are not really interested in whether the points lie on a straight line. Rather we want to know whether they conform to the 45º line, which is the line of equality. This will not be established if we test the null-hypothesis that the correlation is zero.

5. ADDITIONAL REAL DATA EXAMPLES

Reproducibility of ambulatory blood pressure measurements (ABPM)

Ambulatory blood pressure measurements (ABPM) are, notoriously, poorly reproducible. Polynomial curves of ABPM data may be better reproducible than the actual data. Figure 3 gives an example of data.[7] Mean systolic ABPM blood

pressures of 10 untreated patients with mild hypertension ands their sds were recorded twice one week in-between. Figures 2 and 3 give 7th order polynomes of these data. Table 1 shows the results of the reproducibility assessment. Both duplicate sds and ICCs were used. Duplicate sds of means versus zero and versus grand mean were 15.9 and 7.2 mm Hg, while of polynomes they were only 1.86 mm H_g (differences in Duplicate sds significant at a $P<0.001$ level). ICCs of means versus zero and versus grand mean were 0.46 and 0.75, while of polynomes they were 0.986 (differences in levels of correlation significant at a $P<0.001$). Obviously, polynomes of ABPM data of means of populations produce significantly better reproducibility than do the actual data.

Table 1. 24 hr ambulatory blood pressure measurements in a group of 10 patients with untreated mild hypertension tested twice: reproducibility of means of population

	mean values variations vs zero	mean values variations vs grand mean	polynomes
Means (mm Hg) (test 1 / test 2)	153.1 / 155.4	153.1 / 155.4	-
Standard deviation (sd) (mm Hg) (test 1 / test 2)	21.9 / 21.1	15.7 / 13.8	-
95 % CIs [a] (mm Hg) (test 1 / test 2)	139.4-166.8/142.2-168.6	143.3-163.9/146.8-164.0	-
Differences between means (sd) (mm Hg)	-2.4 (22.4)	-2.3 (10.5)	-
P values differences between results tests 1 and 2	0.61	0.51	0.44
Duplicate sds (mm Hg)	15.9	7.2	1.86
Relative Duplicate sds [b] (%)	66	31	7
Intra-class correlations (ICCs)	0.46	0.75	0.986
95 % CIs	0.35-0.55	0.26-0.93	0.972-0.999
Proportion total variance responsible for between-patient variance (%)	46	75	99
95 % CIs (%)	35-55	26-93	97-100

a CIs = confidence intervals.
b Calculated as 100% x [Duplicate sd / (overall mean – 130 mm Hg)].

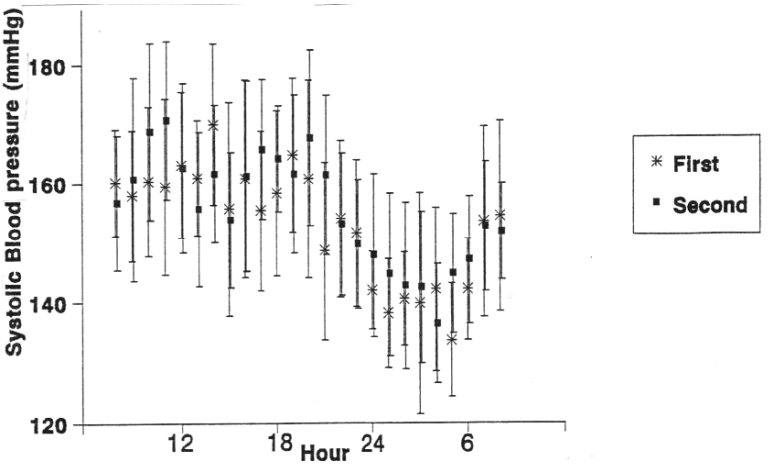


Figure 3. Mean values of ambulatory blood pressure data of 10 untreated patients with mild hypertension and their sds, recorded twice, one week in-between.

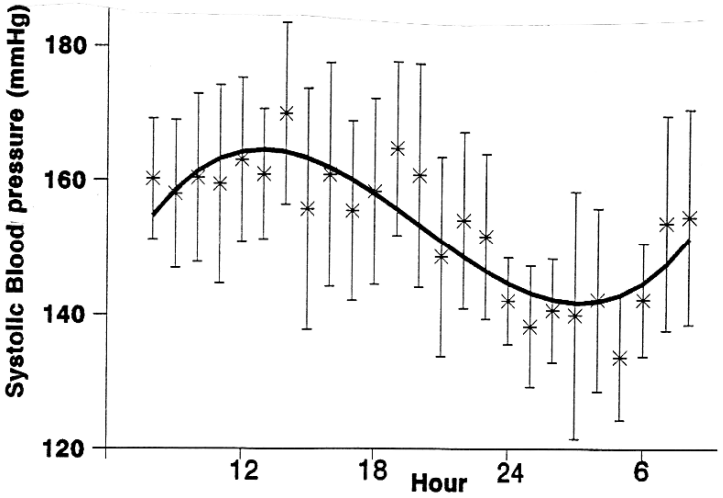


Figure 4. Polynome of corresponding ambulatory blood pressure recording (first one) from Figure 3, reflecting a clear circadian rhythm of systolic blood pressures.

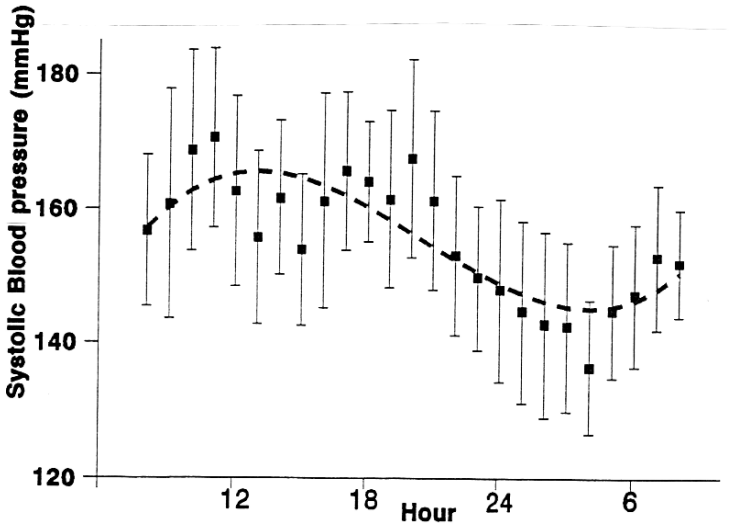


Figure 5. Polynome of corresponding ambulatory blood pressure recording (second one) from Figure 3, again reflecting a clear circadian rhythm of systolic blood pressures.

Two different techniques to measure the presence of hypertension

Two different techniques are used in one group of patients to measure the presence of hypertension, namely (1) ambulatory blood pressure equipments and (2) self-assessed sphygmomanometers. Circumstances are, otherwise, identical.

		Ambulatory equipment		
		yes	no	
Sphygmomanometer	yes	184 (a)	54 (b)	218 (a+b)
	No	14 (c)	63 (d)	77 (c+d)
		198 (a+c)	117 (b+d)	315 (a+b+c+d)

We calculate kappa according to:

expected value for cell (a) = $\frac{184+14}{315} \times 218 = 137$

expected value for cell (d) = $\frac{54+63}{315} \times 218 = 81$

kappa = $\frac{\frac{(218+77)}{315} - \frac{(137+81)}{315}}{1 - \frac{137+81}{315}} = 0.795$

This would mean that we have a substantial level of agreement between the two techniques. However, McNemar's test shows a significant difference at $p < 0.01$

between the two techniques indicating that a systematic difference exists, and that the reproducibility assessment is thus biased. The circumstances are not entirely identical.

6. DISCUSSION

Any research profits from a reproducible challenge test to enhance sensitivity of the trial, and from a good interobserver agreement. The current paper gives some relatively simple methods for assessment. Reproducibility assessments are rarely communicated in research papers and this may contribute to the low reproducibility of clinical trials. We expected that reproducibility testing would, at least, be a standard procedure in clinical chemistry studies where a close to 100% reproducibility is generally required. However, even in a journal like the Journal of the International Federation of Clinical Chemistry and Laboratory Medicine out of 17 original papers communicating novel chemistry methods none communicated reproducibility assessments except for one study.[8] Ironically, this very study reported two incorrect methods for that purpose, namely the assessment of significant differences between repeated measures, and the calculation of Pearson's correlation levels.

A more general explanation for the underreporting of reproducibility assessments in research communications is that the scientific community although devoted to the study of disease management, is little motivated to devote its energies to assessing the reproducibility of the diagnostic procedures required for the very study of disease management. Clinical investigators favor the latter to the former. Also the former gives no clear-cut career path, while the latter more often does so. And there are the injections from the pharmaceutical industry. To counterbalance this is a challenge for governments and university staffs.

We should add that textbooks of medical statistics rarely cover the subject of reproducibility testing: in only one of the 23 currently best sold textbooks for medical statistics the subject is briefly addressed.[9]

We conclude that poor reproducibility of diagnostic criteria / tests is, obviously, a well- recognized but rarely tested problem in clinical research. The current review of simple tests for reproducibility may be of some help to investigators.

7. CONCLUSIONS

Virtually no cardiovascular papers communicate the levels of reproducibility of the diagnostic criteria / tests they use. Poor reproducibility cannot be excluded as a contributing cause for the poor reproducibility of clinical trials. The objective of this chapter was to review simple methods for reproducibility assessment of diagnostic criteria / tests.

Reproducibility of quantitative data can be estimated by (1) duplicate standard deviations, (2) repeatability coefficients, (3) intraclass correlation coefficients. For qualitative data Cohen's kappas are adequate. Incorrect methods include the test for a significant difference between repeated measures, and the calculation of levels of correlation between repeated measures.

Four adequate and two incorrect methods for reproducibility assessment of diagnostic criteria/tests are reviewed. These tests can also be used for more complex data like polynomial models of ambulatory blood pressure measurements. They may be of some help to investigators.

8. REFERENCES

1. Julius S. The ALLHAT study: if you believe in evidence-based medicine. Stick to it. Hypertens 2003; 21: 453-4.
2. Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. Int J Clin Pharmacol Ther 2003; 41: 51-6.
3. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979; 2: 420-8.
4. SPSS 10, SPSS Statistical Software, Chicago, IL, version 2004.
5. Anonymous. Calculating Cohen's kappas. http://www.colorado.edu/geography/gcraft/notes/manerror/html/kappa.html
6. Perloff JK. The clinical recognition of congenital heart disease. Philadelphia, Saunders 1991.
7. Cleophas AF, Zwinderman AH, Cleophas TJ. Reproducibility of polynomes of ambulatory blood pressure measurements. Perfusion 2001; 13: 328-35.
8. Imbert-Bismut F, Messous D, Thibaut V, Myers RB, Piton A, Thabut D, Devers L, Hainque B, Mecardier A, Poynard T. Intra-laboratory analytical variability of biochemical markers of fibrosis and activity and reference ranges in healthy blood donors. Clin Chem Lab Med 2004; 42: 323-33.
9. Petrie A, Sabin C. Assessing agreement. In: Medical statistics at a glance. Blackwell Science, London UK, 2000, page 93.

CHAPTER 33

VALIDATING QUALITATIVE DIAGNOSTIC TESTS

1. INTRODUCTION

Clinical trials of disease management require accurate tests for making a diagnosis/ patient follow-up. Whatever test, screening, laboratory or physical, investigators involved need to know how good it is. The goodness of a diagnostic test is a complex question that is usually estimated according to three criteria: (1) its reproducibility, (2) precision, and (3) validity. Reproducibility is synonymous to reliability, and is, generally, assessed by the size of differences between duplicate measures. Precision of a test is synonymous to the spread in the test results, and can be estimated, e.g., by standard deviations / standard errors. Validity is synonymous to accuracy, and can be defined as a test's ability to show which individuals have the disease in question and which do not. Unlike the first two criteria, the third is hard to quantify, first, because it is generally assessed by two estimators rather than one, namely sensitivity and specificity defined as the chance of a true positive and true negative test respectively. A second problem is, that these two estimators are severely dependent on one another. If one is high, the other is, as a rule, low, vice versa. Due to this mechanism it is difficult to find the most accurate diagnostic test for a given disease. In this chapter we review the current dual approach to accuracy and propose that it be replaced with a new method, called the overall accuracy level. The main advantage of this new method is that it tells you exactly how much information is given by the test under assessment. It, thus, enables you to determine the most accurate qualitative tests for making a diagnosis, and can also be used to determine the most accurate threshold for positive qualitative tests with results on a continuous scale.

2. OVERALL ACCURACY OF A QUALITATIVE DIAGNOSTIC TEST

A test that provides a definitive diagnosis, otherwise called a gold standard test, is 100% accurate. But this test may be too expensive, impractical or simply impossible. Instead, inexpensive but less accurate screening tests, depending on the presence of a marker, are used. Prior to implementation, such tests must be assessed for level of accuracy against the gold standard test. Generally, such tests produce a yes/no result, and are, therefore, called qualitative tests, e.g., the presence of a positive blood culture test, a positive antinuclear antibody test, a positive leuco-esterase urine test, and many more. In order to assess accuracy of such tests, the

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 397–406.

overall accuracy level can be calculated from a representative sample of patients in whom the gold-standard result is known (Table 1).

Table 1. Calculation of sensitivity, specificity, and overall accuracy level of qualitative test from a sample of patients

		Disease	yes(n)	no(n)
Positive test	yes(n)		180 a	20 b
" "	no(n)		30 c	80 d

n= number of patients
a=number of true positive patients
b= false positive patients
c= false negative patient
d= true negative patients

Sensitivity of the above test = a / (a+c) = 180 / 210 = 85.7 %
Specificity = d / (b+d) = 80 / 100 = 80 %.

Overall accuracy level = (a+d) / (a+b+c+d)=260/310 = 83.9 %

The magnitude of the overall accuracy level in the example from Table 1 is 83.9%, which is between that of the sensitivity and specificity, 85.7 and 80 %, but closer to the former than the latter. This is due to the larger number of patients with the disease than those without the disease. Obviously, the overall accuracy level, unlike sensitivity and specificity, adjusts for differences in numbers of patients with and without the disease as generally observed in a representative sample of patients. The overall accuracy level can be interpreted as the amount of information given by the test relative to the gold standard test: if the gold standard test provides 100% of information, the test will provide 83.9% of that information. An overall accuracy of 50% or less indicates that the information is not different from the information provided by mere guessing. Flipping a coin would do the job just as well as does this test. An example of a new test without information is given in Table 2. This new test has a specificity of only 20%, but a sensitivity of 60%, and so the investigators may conclude that it is appropriate to approve this new test, because it provides a correct diagnosis in 60% of the patients who have the disease. However, given the overall accuracy of only 43.8% this diagnostic test does not provide more information than mere guessing or tossing a coin, and should not be approved.

Table 2. Qualitative test providing no more information than mere guessing or tossing a coin

	Disease	yes (n)	no (n)
Positive test	yes	60 a	50 b
Positive test	no	40 c	10 d

n = number of patients
a = number of true positive patients
b = false positive patients
c = false negative patients
d = true negative patients

Sensitivity of the above test= a / (a+c) = 60%
Specificity = d / (b+d) = 20%
Overall accuracy level = (a+d) / (a+b+c+d) = 70/160 = 43.8%

3. PERFECT AND IMPERFECT QUALITATIVE DIAGNOSTIC TESTS

Qualitative diagnostic tests may produce results on a continuous scale, and the results of such tests can be displayed by two Gaussian curves (under the assumption that the data follow normal distributions), rather than simply by a two by two table. Figure 1 is an example of a perfect fit diagnostic test. The two curves show the frequency distribution with on the x-axis the individual patient results and on the y-axis "how often". The total areas under the curve of the two curves represent all of the patients, left graph those without the disease, and right graph those with the disease. The curves do not overlap. The test seems to be a perfect predictor for presence or absence of disease.

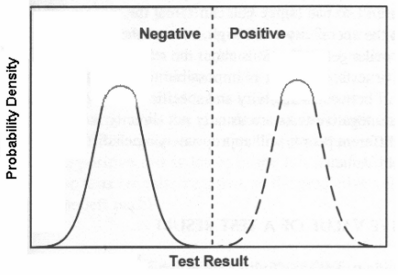


Figure 1. Example of a perfect fit qualitative diagnostic test. The two curves show the frequency distributions with on the x-axis the individual patient results, and on the y-axis "how often". The patients with and without the disease do not overlap.

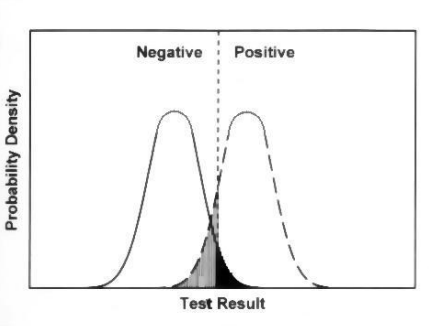


Figure 2. Example of a less than perfect fit qualitative diagnostic test. The 2 curves show the frequency distributions with on the x-axis the individual patient results, and on the y-axis "how often". The patients with and without the disease overlap.

In Figure 2 the situation is less than perfect, the two curves overlap, and, it is not obvious from the graphs where to draw the line between a positive and negative result. The decision made is shown as the vertical line. False positives/negatives are shown in the shaded areas under the curves. The above two examples are simplified, because they assume that in a random sample the total numbers of true positives and true negatives are equal in size, and have the same spread. In practice the numbers of patients with and without disease in a random sample have different sizes and spread, and this should be recognized in the distribution curves, complicating the assessment a little bit (Figure 3).

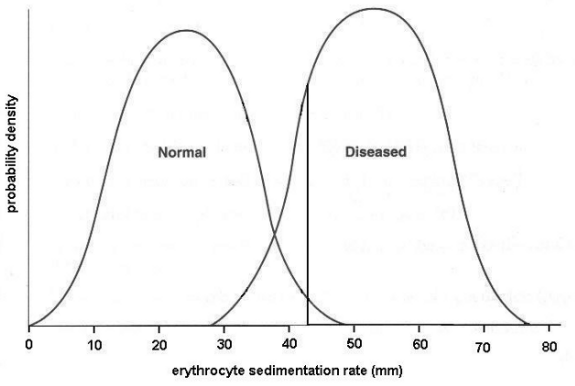


Figure 3. Example of frequency distributions of erythrocyte sedimentation rate values in 200 patients with and 300 patients without pneumonia. On the x-axis are the individual erythrocyte sedimentation rate values of the normals and the diseased patients, and the areas under the curve represent 100% of either of the two groups. It is not obvious from the graphs where to draw the line between a positive and negative test: the decision made is shown by the vertical line.

The left and right graph are calculated from the mean erythrocyte sedimentation rate value and standard deviation of a random sample of patients with and without pneumonia. The areas under the curve represent 100% of either of the two groups. In order to assess accuracy of erythrocyte sedimentation rate as qualitative diagnostic test for pneumonia it is convenient to define a test positive if less than 2.5 % of the true negative patients are negative in the test. Using this 2.5% as a threshold, the results from Figure 3 can now also be displayed in the form of a two by two table (Table 3). Sensitivity, specificity, and overall accuracy are calculated (Table 3).

Table 3. Sensitivity, specificity, and overall accuracy level of qualitative test using the 2.5% threshold for true negative patients (Figure 3)

	Disease	yes(n_1=300)		no(n_2=200)	
Positive test	yes(%)	74%	a	2.5%	b
Positive test	no(%)	26%	c	97.5%	d

n = number of patients
a = number of true positive patients
b = false positive patients
c = false negative patient
d = true negative patients.

Sensitivity of the above test = a / (a+c) = 74 %
Specificity = d / (b+d) = 97.5 %.
Overall accuracy level = $74(n_1 / (n_1 + n_2)) + 97.5 (n_2 / (n_1 + n_2)) = 83.4\%$.

4. DETERMINING THE MOST ACCURATE THRESHOLD FOR POSITIVE QUALITATIVE TESTS

We would like to have a sensitivity and specificity close to 1 (100%), and thus an overall accuracy equally close to 1 (100%). However, in practice most diagnostic tests are far from perfect, and produce false positive and false negative results. We can increase sensitivity by moving the vertical decision line between a positive and negative test (Figure 3) to the left, and we can increase specificity by moving it in the opposite direction. Moving the above threshold further to the right would be appropriate, e.g., for an incurable deadly disease. You want to avoid false positives (cell b), meaning telling a healthy person he/she will die soon, while false negatives (cell c) aren't so bad since you can't treat the disease anyway. If, instead the test would serve for a disease fatal if untreated but completely treatable, it should provide a sensitivity better than 74%, even at the expense of a lower specificity.

False-negative would be awful, as it means missing a case of a treatable fatal disease. For that purpose the threshold of such a test is set far more to the left (Figure 4).

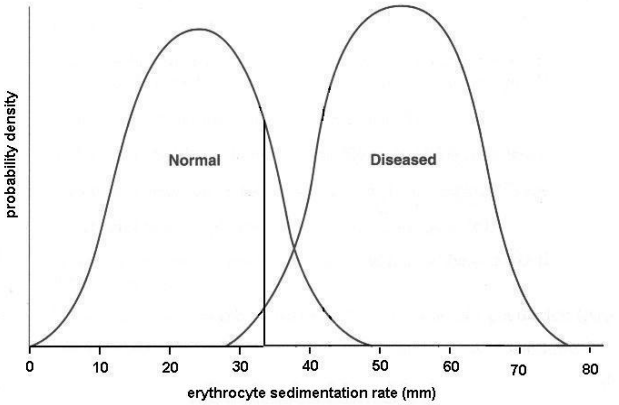


Figure 4. Example of the frequency distributions of erythrocyte sedimentation rate values in 200 patients with and 300 patients without pneumonia. On the x-axis are the individual erythrocyte sedimentation rate values of the normals and the diseased patients, and the areas under the curve represent 100% of either of the two groups. It is not obvious from the graphs where to draw the line between a positive and negative test: the decision made is shown as the vertical line.

Sensitivity, specificity, and overall accuracy level can now be calculated (Table 4).

Sensitivity of the above test = a / (a+c) = 97.5 %
Specificity = d / (b+d) = 77 %
Overall accuracy level = $97.5(n_1 / (n_1 + n_2)) + 77 (n_2 / (n_1 + n_2))$ = 89.3%

Table 4. Calculation of sensitivity, specificity, and overall accuracy level of a qualitative test where the threshold is set according to Figure 4

	Disease	yes(n_1=300)		no(n_2=200)	
Positive test	yes(%)	97.5%	a	23%	b
" "	no(%)	2.5%	c	77%	d

n_x = number of patients
a = % true positive patients
b = % false positive patient
c = % false negative patient
d = % true negative patients

There are, of course, many diseases that do not belong to one of the two extremes described above. Also, there may be additional arguments for choosing a particular threshold. E.g., in non-mortality trials false negative tests, generally, carry the risk of enhanced morbidity, such as vision loss due to persistent glaucoma, hearing loss due to recurrent otitis etc. However, such risks may be small if repeat tests are performed in time. Also, false positive tests create here patient anxiety and costs. In situations like this, false positive tests are considered as important as false negative. Therefore, we might as well search for the threshold providing the best overall accuracy from our test. This is usually done by considering several cut-off points that give a unique pair of values for sensitivity and specificity, thus comparing the probabilities of a positive test in those with and those without the disease. A curve with "1-specificity" (= proportion of false positive tests) on the x-axis and sensitivity (= proportion of true positive tests) on the y-axis facilitates to choose cut-off levels with relatively high sensitivity/specificity. The continuous curve of Figure 5, otherwise called a ROC (receiver operating characteristic) curve, is an example.

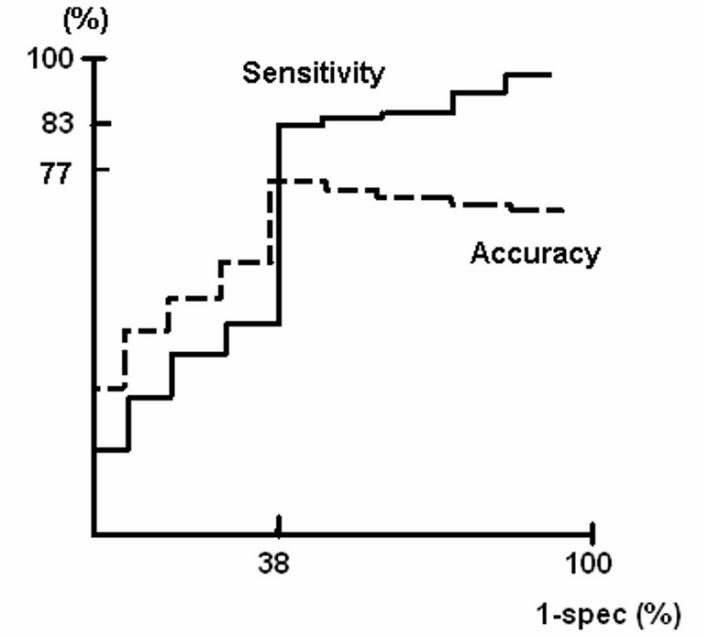


Figure 5. The ROC (receiver operating characteristic) curve (continuous curve) of erythrocyte sedimentation rate values of patients with pneumonia plots the sensitivity values (true positives) against the "1-specificity" values (false positives).

The accuracy "ROC" curve (interrupted curve) plots the overall accuracy values against the "1-specificity" values (false positives).

It shows the relationship between sensitivities and specificities of the erythrocyte sedimentation rate as a diagnostic test for the presence of pneumonia. The curve suggests that a relatively high sensitivity/specificity is obtained for the 83% sensitivity/ 38% "1-specificity". However, in many ROC curves more than a single cut-off value with relatively high sensitivity/specificity are observed, and it may, therefore, be difficult to choose the most accurate cut-off level from such curves. Also, ROC curves use sensitivity and specificity only, which means that they do not account for differences between the numbers of patients with and without the disease. These problems can be prevented by plotting, instead of the sensitivity, the overall accuracy level against "1-specificity". This is shown by the interrupted curve of Figure 5. This accuracy "ROC" curve will unequivocally identify the cut-off threshold with the single best overall accuracy level.
ROC curves are only briefly addressed in this text. Details are beyond the scope of this chapter, but some advantages of accuracy "ROC" curves compared to the classic ROC curves are mentioned. In addition to ROC curves, accuracy is sometimes assessed by measure of concordance, the optimism corrected c-statistic (Young 1948).[2] It is identical to the AUC (area under the curve) of the ROC curve, and varies between 0.5 and 1.0. The larger the AUC, the better the accuracy.

5. DISCUSSION

Another approach to accuracy of diagnostic tests are the positive and negative predictive values and likelihood ratios, the calculation of which is shown in Table 5.

Table 5. The calculation of positive and negative predictive values, and of likelihood ratios

	Disease	yes(n)	no(n)
Positive test	yes	a	b
" "	no	c	d

n = number of patients
positive predictive value = a / (a+b)
negative predictive value = d / (c+d)
likelihood ratio for positive result = a / (a+c) / d / (b+d)

Just like the overall accuracy level, these estimators adjust for numbers of differences in patients with and without the disease, but they do not answer what proportion of patients has a correct test.

Riegelman[1], recently, proposed another method for assessing accuracy of a qualitative diagnostic test, which he called the discriminant ability, defined as

(sensitivity+specificity)/2.

Although this method avoids the dual approach to accuracy, it wrongly assumes equal importance and equal prevalence of sensitivity and specificity, and does neither answer what proportion of the patients has a correct test.

We should add that sensitivity, specificity and overall accuracy level are usually expressed as percentages. As with all estimates in clinical trials, we should calculate confidence intervals of these estimates in order to quantify the level of uncertainty involved in our results (see Chapter 34).

The advantage of the overall accuracy approach described in this chapter compared to the dual sensitivity/specificity approach is that it enables to determine not only the most accurate qualitative tests for making given diagnoses, but also the most accurate thresholds for positive qualitative tests with results on a continuous scale. The method is less adequate for the assessment of diagnostic tests for extreme disease like incurable deadly diseases and treatable but untreated deadly diseases for which diagnostic tests with either optimal sensitivity or optimal specificity are required.

For determining the most accurate threshold for a qualitative test we recommend to replace a ROC curve with an accuracy "ROC" curve, because the latter unlike the former accounts for possible differences in a random sample between the numbers of patients with and without the disease.

The overall accuracy level has four advantages compared to the sensitivity / specificity levels. It (1) adjusts for differences between numbers of patients with and without the disease, (2) is able to readily identify tests that give no information at all, (3) provides the amount of information given by the test relative to the gold standard test, (4) enables to draw ROC curves adjusted for the differences between numbers of patients with and without the disease.

6. CONCLUSIONS

Clinical trials of disease management require accurate tests for making a diagnosis / patient follow-up. Currently, accuracy of qualitative diagnostic tests is hard to quantify, because it is generally assessed by two estimators, sensitivity and specificity, that are severely dependent on one another. If one estimator is high, the other is, as a rule, low.

The objective of this chapter was to review the current dual approach to accuracy, and to propose that it be replaced with a new method, called the overall accuracy level.

The overall accuracy level is defined as the proportion of test results that are correct. Usage of this level, unlike sensitivity and specificity levels, enables (1) to adjust for differences between numbers of patients with and without the disease, (2) to readily identify tests that give no information at all, (3) to provide the entire

amount of information given by the test relative to the gold standard test, (4) to draw receiver operating characteristic (ROC) curves adjusted for the differences between numbers of patients with and without the disease. The method is less adequate for the assessment of qualitative diagnostic tests for extreme diseases like incurable deadly diseases and treatable but untreated deadly diseases for which diagnostic tests with either optimal sensitivity or optimal specificity are required.
Due to the dual sensitivity/specificity approach to accuracy of qualitative diagnostic tests it is, currently, difficult to find the most accurate diagnostic test for a given disease. The overall accuracy level is more appropriate to that aim.

7. REFERENCES

1. Riegelman RK. Studying a study and testing a test. Lippincott Williams & Wilkins, Philadelphia, PA, 2005.
2. Anonymous. Chapter 8: statistical models for prognostication. In : Interactive Textbook. http:// symptomresearch.nih.gov.chapter_8.

CHAPTER 34

UNCERTAINTY OF QUALITATIVE DIAGNOSTIC TESTS

1. INTRODUCTION

In cardiovascular research gold standard tests for making a diagnosis are often laborious and sometimes impossible. Instead, simple and non-invasive tests are often used. A problem is that these tests have limited sensitivities and specificities. Levels around 50% means that no more information is given than flipping a coin. Levels substantially higher than 50% are commonly accepted as documented proof, that the diagnostic test is valid. However, sensitivity / specificity are estimates from experimental samples, and scientific rigor recommends that with experimental sampling amounts of uncertainty be included. Although the STARD (Standards for Reporting Diagnostic Accuracy) working party recently advised "to include in the estimates of diagnostic accuracy adequate measures of uncertainty, e.g., 95%-confidence intervals"[1], so far uncertainty is virtually never assessed in sensitivity / specificity evaluations of cardiovascular diagnostic tests. This is a pity, because calculated levels of uncertainty can be used for statistically testing whether the sensitivity / specificity is significantly larger than 50%. The present study uses examples to describe (1) simple methods for calculating standard errors and 95% confidence intervals, and (2) how they can be employed for statistical testing whether the new test is valid. We do hope that this paper will stimulate cardiovascular investigators to more often assess the uncertainty of the diagnostic tests they apply.

2. EXAMPLE 1

Two hundred patients are evaluated the determine the sensitivity / specificity of B-type Natriuretic Peptide (BNP) for making a diagnosis of heart failure.

		Heart failure (n)	
		Yes	No
Result diagnostic test	positive	70 (a)	35 (b)
	negative	30 (c)	65 (d)

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 407–413.

The sensitivity (a / (a+c)) and specificity (d / (b+d)) are calculated to be 0.70 and 0.65 respectively (70 and 65 %). In order for these estimates to be significantly larger than 50% their 95% confidence interval should not cross the 50% boundary. The standard errors are calculated according to the equations given in Appendix 1. For sensitivity the standard error is 0.0458, for specificity 0.0477. Under the assumption of Gaussian curve distributions in the data the 95% confidence intervals of the sensitivity and specificity can be calculated using the equations

95% confidence interval of the sensitivity = 0.70 ± 1.96 x 0.0458
" " " specificity = 0.65 ± 1.96 x 0.0477.

This means that the 95% confidence interval of the sensitivity is between 61% and 79 %, for specificity it is between 56% and 74%. These results do not cross the 50% boundary and fall, thus, entirely within the boundary of validity. The diagnostic test can be accepted as being valid.

3. EXAMPLE 2

Dimer tests have been widely used as screening tests for lung embolias.

		Lung embolia (n)	
		Yes	No
Dimer test	positive	2 (a)	18 (b)
	negative	1 (c)	182 (d)

The sensitivity (a / (a+c)) and specificity (d / (b+d)) are calculated to be 0.666 and 0.911 respectively (67 and 91 %). In order for these estimates to be significantly larger than 50% the 95% confidence interval of them should again not cross the 50% boundary.
The standard errors as calculated according to the equations given in Appendix 1, are for sensitivity 3.672, for specificity 0.286. Under the assumption of Gaussian curve distributions the 95% confidence intervals of the sensitivity and specificity are calculated using the equations

95% confidence interval of the sensitivity = 0.67 ± 1.96 x 3.672
" " " specificity = 0.91 ± 1.96 x 0.286.

The 95 % confidence interval of the sensitivity is between -5.4 and + 7.8. The 95% confidence interval of the specificity can be similarly calculated, and is between 0.35 and 1.47. These intervals are very wide and do not at all fall within the boundaries of 0.5-1.0 (50 – 100 %). Validity of this test is, therefore, not really demonstrated. The appropriate conclusion of this evaluation should be: based on

this evaluation the diagnostic cannot be accepted as being valid in spite of a sensitivity and specificity of respectively 67 and 91%.

4. EXAMPLE 3

A disadvantage of the sensitivity / specificity approach to validation is that it is dual and that the two estimates are severely dependent on one another. Instead, overall-validity is sometimes used. It is defined as the diagnostic test's ability to show which individuals have a true test either positive or negative ((a+d)/(a+b+c+d) where the letters indicate the numbers of patients in the 4 cells as demonstrated above).
As an example, for approval of C-reactive protein as a marker for a cardiovascular event a boundary of overall-validity is specified in the study protocol as being at least 85%. The 95% confidence interval of the overall validity level can be calculated from the data. If the confidence interval falls entirely within the specified boundary, overall validity is demonstrated.
the results are given underneath:

sensitivity = 80 % with a standard error = 2%,
specificity = 90% with a standard error =1%,
prevalence =10% with a standard error =3%.

With this information we can calculate the overall-validity using the method described in Appendix 2. The overall-validity equals 0.89 (89%), while its squared standard error, otherwise called variance, equals 0.000337.
The standard error of the overall-validity is, thus, the square root of its variance, and equals 0.01836 (1.836%). An overall-validity of 89% with a standard error of 1.836 % means that the 95% confidence interval is between 0.89 - (1.96 x 0.01836) and 0.89 + (1.96 x 0.01836), and is thus between 85.4 and 92.6 %. This interval falls entirely between the specified interval of validity of at least 85%. The overall-validity of this diagnostic test has been demonstrated.

5. EXAMPLE 4

A methionine loading test is applied to assess cystathione-beta-synthase deficiency, an inborn error of metabolism causing homocystinuria. The gold standard test is the measurement of the intracellular lacking enzyme, a laborious method.

		cystathione synthase deficiency (n)	
		Yes	No
Methionine loading test	Yes	18 (a)	17 (b)
	No	2 (c)	31 (d)

In this evaluation the sensitivity and specificity were adequate (0.90 and 0.65 respectively). However, in the protocol the investigators pre-specified their boundary of overall-validity between 0.5 and 1.0 (50 and 100%). For assessment of uncertainty and statistical testing the method of Appendix 2 was applied. Overall-validity equalled 0.7205, its standard error 0.1355. The 95% confidence intervals of the overall-validity were calculated to be between 0.7205 ± (1.96 x 0.1355) and is, thus, between 0.45 and 0.99. This confidence interval is wide, and does not entirely fall within the pre-specified boundaries. According to the presented assessment the validity of this test could not be confirmed. With larger samples this validation-procedure might have been more successful.

6. DISCUSSION

The accuracy of cardiovascular diagnostic tests is often assessed by sensitivity, specificity, and sometimes by overall-validity, but the precision of these point estimates is rarely taken into account. Low precision means that the 95% confidence interval of them is wide, and, thus, that the diagnostic tests can not be reliably used for making predictions. In this paper it is shown that, by calculating the standard error of the diagnostic test, its precision can be assessed. From the standard errors 95% confidence can be calculated. If the 95% confidence intervals of the standard error fall entirely within pre-defined boundaries, the diagnostic test can be accepted as being valid. If not, the test should be rejected.
We should add that the sample size is, of course, a major determinant of the confidence intervals. For example, according to the above methods the 95% confidence intervals of the proportion of true positives (sensitivity) with:

n = 10 is between 0.410 – 0.990
n = 100 is between 0.610 – 0.790
n = 1000 is between 0.671 – 0.729.

The validation samples should, therefore, largely match the sample sizes of the future clinical trials using the diagnostic test under study. If the size of your validation sample n = 100, then this diagnostic test is probably not adequately sensitive/specific for a clinical trial including a sample size of n = 10. In contrast, if clinical trials include many more patients than included in the validity assessment of their diagnostic tests, then the confidence intervals are underestimated, and the diagnostic test will perform even better than predicted by the calculated confidence intervals.
We conclude that adequate diagnostic tests are vital for the multitude of cardiovascular intervention studies of new therapies. For validation of diagnostic tests sensitivity / specificity / overall-validity are calculated. This practice is incomplete, because the number of true positives and true negatives in your assessment are estimates from experimental samples, and scientific rigor requires that with any estimate in clinical research standard errors / confidence intervals have to be included in order to quantify the level of uncertainty in your data. We

provide simple methods for such purpose, and do hope that they be implied in future validation studies of cardiovascular diagnostic tests.

7. CONCLUSION

In clinical research simple and non-invasive tests are often used instead of the gold standard tests for making diagnoses. Because the sensitivity / specificity of the simple tests are limited, their magnitude is routinely accounted in validation procedures. These measures of validity are estimates from experimental samples, and their precision, otherwise called certainty, is rarely assessed. This chapter gives simple methods for establishing their uncertainty.

As with other estimates in clinical research the standard errors of the sensitivity and specificity can be calculated in order to quantify their uncertainty. From these standard errors confidence intervals can be calculated. In the study protocol validation boundaries of the confidence intervals should be pre-specified. Only, if the confidence intervals fall entirely within these validation boundaries, validity is demonstrated. We recommend that the lower level of the validation boundaries should never be set below 50%, because a sensitivity and specificity close to 50% gives no more information than tossing a coin.

An effort should be made to assess uncertainty of the sensitivity and specificity of diagnostic tests before accepting them for general use. Simple methods for that purpose are given.

8. REFERENCES

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig JG, Moher D, Rennie D, De Vet HC, for the STARD steering group. Education and Debate. Towards Complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326: 41-4.
2. Berger JO, Bernerdo J. Estimating a product of normal means: Bayesian analysis with some priors. J Am Stat Assoc 1989; 84: 200-7.
3. Anonymous. Delta-method. Http://en.wikidepia.org/wiki/Delta_method

APPENDIX 1

For the calculation of the standard errors (SEs) of sensitivity, specificity and overall-validity we make use of the Gaussian curve assumption in the data.

		Definitive diagnosis (n)	
		Yes	No
Result diagnostic test	Yes	a	b
	No	c	d

Sensitivity = a / (a+c) = proportion true positives
Specificity = d / (b+d) = proportion true negatives
1-specificity = b / (b+d)
Proportion of patients with a definitive diagnosis = (a+c) / (a+b+c+d)
Overall validity = (a+d) / (a+b+c+d)

In order to make predictions from these estimates of validity their standard deviations / errors are required. The standard deviation / error (SD/ SE) of a proportion can be calculated.

SD =	$\sqrt{}$ p(1-p) where p = proportion.
SE =	$\sqrt{}$ [p(1-p) / n] where n = sample size

where p equals a/(a+c) for the sensitivity. Using the above equations de standard errors can be readily obtained.

$SE_{\text{sensitivity}}$ =	$\sqrt{}$ ac / $(a+c)^3$
$SE_{\text{specificity}}$ =	$\sqrt{}$ db/ $(d+b)^3$
$SE_{\text{1-specificity}}$ =	$\sqrt{}$ db / $(d+b)^3$
$SE_{\text{proportion of patients with a definitive diagnosis}}$ =	$\sqrt{}$ (a+b)(c+d) / $(a+b+c+d)^3$

APPENDIX 2

The equation of the SE of the overall-validity is less straightforward, but can be obtained using the Bayes' rule[2] and the delta method.[3] The calculations are given for the purpose of completeness (Var = variance = square root of the standard error; prevalence = proportion of patients with a definitive diagnosis).

Overall-validity =	sensitivity x prevalence + specificity x (1-prevalence)

In order to calculate the standard error (SE), we make use of the equation (Var = variance, Cov = covariance)

Var(X+Y) =	Var(X) + Var(Y) + 2 Cov(X,Y)

If X = sensitivity x prevalence, and Y = specificity x (1-prevalence), then the equations can be combined to obtain an equation for the variance of the overall-validity (sens = sensitivity, spec = specificity, prev = prevalence)

$Var_{\text{overall validity}}$ =	$Var_{\text{sens x prev}}$ + $Var_{\text{spec x (1-prev)}}$ + 2 $Cov_{\text{sens x prev, spec x (1-prev)}}$

The variance of X + Y may according to the delta-method [3] be approached from:

$Var(X + Y) =$	$Y^2\ Var(X) + X^2\ Var(Y)$

By combining the equations we will end up finding:

$Var_{\text{overall validity}} =$	$prev^2 \times Var_{\text{sens}} + (1\text{-}prev)^2 \times Var_{1\text{-spec}} + (sens\text{-}spec)^2 \times Var_{\text{prev}}$

The delta-method describes the variance of natural logarithm (ln) (X) as Var(ln(x)) = Var(x) / x^2 . The approach is sufficiently accurate if the standard errors of prevalence, sensitivity and specificity are small, which is true if samples are not too small.

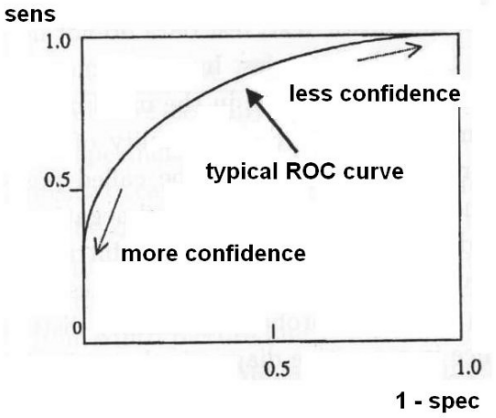
CHAPTER 35

META-ANALYSIS OF DIAGNOSTIC ACCURACY STUDIES

1. INTRODUCTION

An intuitive approach to meta-analysis of qualitative diagnostic tests is to separately pool sensitivities and specificities of separate studies by standard methods. Reported sensitivities and specificities are, however, negatively correlated, and, in addition, in a curvilinear manner as commonly shown in receiver operated characteristic (ROC) curves (Figure 1). It has been demonstrated, that it is appropriate to take this negative and curvilinear correlation into account in the meta-analysis.[1]

In the past few years many novel diagnostic methods have been developed, including multi-slice computer tomography, magnetic resonance, positive emission tomography and many more methods. All of these methods need validation, before they can be meaningfully used in research and clinical practice. This chapter reviews methods for pooling the meta-data of qualitative diagnostic tests, while accounting for the negative correlation between sensitivity and specificity. We should emphasize that such meta-data should, of course, also be assessed for the usual pitfalls of meta-analyses as described in Chapter 33. The STARD[2,3] statement now used by the major journals contains a flowchart, and a checklist helps to ensure complete reporting of the design and results of the diagnostic accuracy studies, and will facilitate future meta-analyses.



T.J. Cleophas et al., Statistics Applied to Clinical Trials, 415–421.

Figure 1. Example of summary receiver operated characteristic (ROC) curve, the true positives(sens) are drawn against the false positives (1-spec) using the results of multiple studies.

2. DIAGNOSTIC ODDS RATIOS (DORS)

For a particular study the diagnostic data can be summarized in the well known two-by-two table:

		Reference test	
		Positive	*Negative*
Diagnostic Test	*Positive*	True Positive Rate (TPR=sensitivity)	1-TNR
	Negative	1-TPR	True Negative Rate (TNR=specificity)

Compared to the reference test which is used here as a gold-standard, the accuracy of the diagnostic test is summarized by two statistics: the proportion of patients with a positive reference test in who the diagnostic test is positive, and the proportion with a negative reference test in who the diagnostic test is negative. The two statistics are well known as the true-positive-rate (TPR) or sensitivity, and the true-negative-rate (TNR) or specificity and are often used to draw summary ROC curves (Figure 1).
Instead of the dual approach of sensitivity and specificity, accuracy can also be summarized by the diagnostic odds ratio (DOR):

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

The DOR is an interesting term, since it compares the odds of true positive patients with that of false positives, and, in a way, summarizes the overall accuracy of a diagnostic test. A problem is, that, like any odds ratio, it does not follow a Gaussian distribution, and a logarithmic transformation is required. This is why it is used in the so-called summary receiver operated characteristic (sROC) -approach of Moses and Littenberg.[1] This approach entails the natural logarithm of the diagnostic odds ratio [ln(DOR)] and the statistic S:

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

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

where TPR and FPR are the true and false positive rates. A linear regression analysis model according to ln(DOR) = α + β.S is used to fit the data, and is often successful for that purpose (α intercept and β regression coefficient). The ln(DOR) is, thus, an overall indicator for diagnostic accuracy, and indicates how more often a positive test result will occur among patients with the condition of interest compared to patients without. S relates to the test threshold. It has a value of 0.00 in a study where sensitivity equals specificity. S is positive in studies where sensitivity is higher than specificity, and S will be negative if specificity is higher. An example as previously used by the authors' working party is taken (Table 1).[4,5] Three different regression lines were fitted. The intercepts (α) and slopes (β) are in Table 2.

Table 1. Example of meta-analysis of 44 qualitative diagnostic studies[4]

Study No.	diagnostic modality	tp	fp	fn	tn
1.Grumbine 1981	1	0	1	6	17
2.Walsh 1981	1	12	3	3	7
3.Brenner 1982	1	4	1	2	13
4.Villasanta 1983	1	10	4	3	25
5.Van Engelshoven 1984	1	3	1	4	12
6.Bandy 1985	1	9	3	3	29
7.Vas 1985	1	20	4	8	31
8.King 1986	1	17	5	7	21
9.Feigen 1987	1	2	0	9	32
10.Camilian 1988	1	3	1	9	38
11.Janus 1989	1	1	1	2	18
12.Matsukuma 1989	1	5	2	2	61
13.Heller 1990	1	21	8	40	184
14.Kim 1990	1	4	3	9	42
15.Ho 1992	1	0	0	5	15
16.Kim 1993	1	7	11	22	158
17.Subak 1995	1	3	3	2	29
18.Kindenmann 1979	2	19	1	10	81
19.Lecart 1971	2	8	9	2	13
20.Piver 1971	2	41	1	12	49
21.Piver 1973	2	5	1	2	18
22.Kolbenstvedt 1975	2	45	58	32	165

23.Leman 1975	2	8	6	2	32
24.Brown 1979	2	5	8	1	7
25.Lagasse 1979	2	15	17	11	52
26.Kjorstad 1980	2	16	11	8	24
27.Ashraf 1982	2	4	8	2	25
28.De Muylder 1984	2	4	12	10	70
29.Smales 1986	2	10	4	4	55
30.Feigen 1087	2	2	5	6	23
31.Swart 1989	2	7	10	7	30
32.Hellert 1990	2	4	50	12	135
33.Lafianza 1990	2	8	3	1	37
34.Stellato 1992	2	4	3	0	14
35.Hricak 1988	3	9	2	2	41
36.Greco 1989	3	3	6	5	32
37.Janus 1989	3	3	2	1	16
38.Kim 1990	3	3	1	12	44
39.Ho 1992	3	0	0	5	15
40.Kim 1993	3	7	2	22	167
41.Hawnaur 1994	3	12	4	4	29
42.Kim 1994	3	23	5	14	230
43.Subak 1995	3	8	5	5	53
44.Heuck 1997	3	16	2	2	22

Diagnostic modality: lymphangiography (1), computed tomography (2), and magnetic resonance (3); tp = true positive, fp = false positive, fn = false negative, tn = true negative.

DOR are calculated by back-transformation of the logit terms. The interpretation of the intercept and the slopes of the linear regression line is not straightforward When the diagnostics odds ratio (DOR) does not depend on the threshold S (regression coefficient ≈ 0), then the intercept would provide a summary estimate of the DOR. When the DOR varies with S, then the regression coefficient has no direct interpretation, but has a substantial influence on the shape of the ROC curve. Usually, DOR at mean of S or the Q-point (the point where the descending diagonal of the ROC graph cuts the ROC curve) are taken as overall estimates. However, a summary estimate is not available. An advantage of the regression models is, that it is easy to extend the model with covariates, representing differences between studies in design. But this advantage is limited, because these covariates are supposed to affect sensitivity and specificity in a similar manner, which need not be the case.

Table 2. Intercepts and slopes of the linear regression lines of the DORs of three diagnostic modalities from Table 1

diagnostic modality	intercept (SE)	regression(SE) coefficient	DOR at mean S (95% CI)	Q-point (95% CI)
LAG	2.1 (0.3)	-0.4 (0.2)	16.0 (8.4-30.7)	0.7 (0.7-0.8)
CT	2.8 (0.4)	0.2 (0.1)	10.9 (6.5-18.3)	0.8 (0.7-0.9)
MRI	3.5 (0.6)	0.3 (0.2)	20.3 (10.3-39.7)	0.9 (0.8-0.9)
p-value				
LAG vs CT			0.36	0.15
LAG vs MRI			0.62	0.01
CT vs MRI			0.15	0.34

LAG = lymphangiography; CT = computed tomography; MRI = magnetic resonance.

3. BIVARIATE MODEL

A summary estimate of DOR is provided by the bivariate model.[6] It is adjusted for interaction between sensitivity and specificity, and based on loglinear mixed effects modeling. More information about mixed effects models will be given in Chapter 40. Briefly, the following reasoning is applied. We assume, that the sensitivities from the individual studies after logit transformation follow a Gaussian curve around a mean value. The same applies to the specificities. The combination of two Gaussian-like distributions and their logit transformations lead to bivariate normal distributions, while interaction between the two distributions is taken into account. This bivariate model can be analyzed using SAS Proc Mixed Statistical Software[7] after the following commands:

```
*/proc mixed data=bj_meta method=reml cl
*/model logit
*/random dis non_dis
*/repeated / group=rec
*/contrast ‘CT sens vs LAG sens’;
  contrast ‘CT sens vs MRI sens’ ;
  contrast ‘LAG sens vs MRI sens’.
```

Similarly, differences in specificities and DORs are calculated and tested.

Table 3. Summary estimates for sensitivity, specificity, and diagnostic odds ratio as calculated by the bivariate model for the studies from Table 1

diagnostic	mean (95% CI)	mean (95% CI)	Mean (95% CI)

modality	sensitivity	specificity	DOR
LAG	0.67 (0.57-0.76)	0.80 (0.73-0.85)	8.13 (5.16-12.82)
CT	0.49 (0.37-0.61)	0.92 (0.88-0.95)	11.34 (6.66-19.30)
MRI	0.56 (0.41-0.70)	0.94 (0.90-0.97)	21.42 (10.81-42.45)
p-value			
LAG vs CT	0.023	0.0002	0.35
LAG vs MRI	0.21	0.0001	0.021
CT vs MRI	0.47	0.34	0.15

LAG = lymphangiography; CT = computed tomography; MRI = magnetic resonance.

Table 3 gives the summary estimates for sensitivity, specificity, and diagnostic odds ratio as calculated by the bivariate model for the studies from Table 1. The bivariate model has advantages compared to the above DOR model. First, it enables to provide summary estimates of sensitivities, specificities and DORs adjusted for interaction between sensitivities and specificities. Second, using the parameters of the bivariate distribution, we can calculate rectangular confidence intervals around the mean values of the logit sensitivity and specificity. Like the DOR model, explanatory variables can be added, but, unlike the DOR model, this leads to separate effects on sensitivity and specificity, rather than a net effect on the DOR scale.

4. CONCLUSIONS

Reported sensitivities and specificities of different studies are not only negatively correlated, but also in a curvilinear manner. It is appropriate to take this negative curvilinear correlation into account in meta-analyses.
Diagnostic odds ratios and a bivariate mixed effects model can be used for data pooling. Prior to pooling it is also appropriate to assess the meta-data for the usual pitfalls of meta-analysis including publication bias, clinical heterogeneity, and lack of robustness as reviewed in Chapter 33.

5. REFERENCES

1.Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993; 12: 1293–316.
2. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate

reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem. 2003; 49: 1-6
3. STARD Statement. http://www.consortstatement.org/stardstatement.htm
4. Scheidler J, HricackH, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. JAMA 1997; 278: 1096-101.
5. Reitsma J, Glas A, Rutjes A, Scholten R, Bossuyt P, Zwinderman A bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005: 58: 982–90.
6. Van Houwelingen HC, Zwinderman AH, Stijnen T. A bivariate approach to meta-analysis. Stat Med 1993; 12: 2273–84.
7. SAS Statistical Software. www.SAS.com

CHAPTER 36

VALIDATING QUANTITATIVE DIAGNOSTIC TESTS

1. INTRODUCTION

Clinical research is impossible without valid diagnostic tests. The methods for validating *qualitative* diagnostic tests include sensitivity / specificity assessments and ROC (receiver operated characteristic) curves, and are generally accepted. [1-4] In contrast, the methods for validating *quantitative* diagnostic tests have not been agreed upon by the scientific community.[4] This paper, using real data examples, reviews the advantages and disadvantages of various methods that could be used for that purpose.

2. LINEAR REGRESSION TESTING A SIGNIFICANT CORRELATION BETWEEN THE NEW TEST AND THE CONTROL TEST

Regression methods are often used for that purpose, particularly, linear regression using a significant correlation as criterion for validation. In Figure 1 an example is

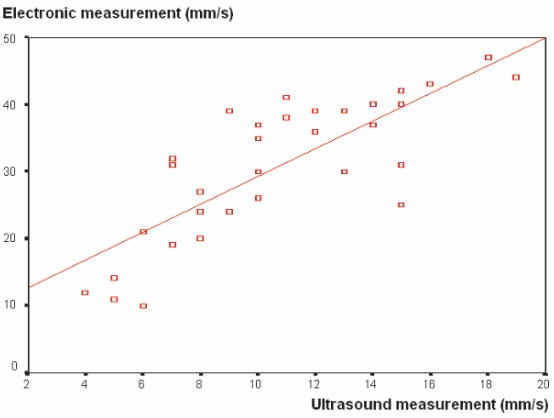


Figure 1. Validity assessment with a linear regression model. The regression equation is given by y = a + b x = 8.647 + 2.065 x (a = intercept, b = regression coefficient, p <0.0001). The x-axis-data, ultrasound estimates, are a very significant predictor of the y-axis-data, the electromagnetic measurements.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 423–431.

However, the prediction, despite the high level of statistical significance, is very imprecise. E.g., if x = 6, then y may be 10 or 21, if x = 7, y may be 19, 31 or 32.

given. A positive correlation seems to exist between the new-test- and control-test-data given by the x-axis-data and the y-axis-data. We can draw a best fit regression line according to the equation

$$y = a + b\,x$$

For every x-axis-datum this line provides the best predictable y-axis-datum. The b-value is the regression coefficient (= direction coefficient), "a" the intercept, which is the place where the line crosses the y-axis. The values "a" and "b" from the equation $y = a + b\,x$ can be calculated:

$$b = \text{regression coefficient} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

This equation is often described in a condensed way as

b = SP xy / SS x (SP xy = sum of products of x- and y-data, SS x = sum of squared x-data)

a = intercept = $\bar{y} - b\,\bar{x}$.

Another important term for regression analyses is the r-value, the correlation coefficient,

$$r = \text{correlation coefficient} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

The term r gives the measure for strength of association between the x-data and the y-data. The stronger the association, the better the x-data predict the y-data. It varies from –1 to +1, r = 0 means no association at all, r = – 1 or = + 1 means 100% association (we can predict the y-values from the given x-values with 100% certainty).
The term r^2 is often more convenient, because it varies from 0 to +1. The r^2 – value is used as a measure of the percentage certainty that has been obtained by the linear regression model. For example, an r^2 – value of 0.36 means that we can predict the y-data with 36% certainty, if we know the corresponding x-data.
For statistical testing of linear regression lines we test with the Student's t-test whether the b-value is significantly larger than zero or with analysis of variance whether the r^2 –value is significantly larger than zero. These tests are laborious,

and, therefore, currently routinely performed by statistical software. For example, SPSS statistical software requires after entering the data the commands: statistics; regression; linear. In the given example (Figure 1) the b-value is calculated to be 2.065 with a standard error of 0.276 and a t-value of 7.491, meaning that it is , indeed, significantly larger than zero at $p<0.0001$, and that there is, thus, a strong significant association between the new-test-data and the control-test-data. Also the r^2 – value of 0.63 as calculated is significantly larger than 0 at $p<0.0001$. Both results, thus, indicate that a significant association exists between the x-data and the corresponding y-data. This means that the data are significantly closer to the regression line than could happen by chance. However, it does not mean that they are all situated exactly on the regression line. As can be observed in the Figure 1, if, for example, x = 6 then y may be 10 or 21, if x = 7 then y may be 19, 31 or 32. Actually, given the r^2 – value of 0.63, we may conclude that any particular x-datum can predict the corresponding y-datum only by 63%, while 37% remains uncertain. This percentage of uncertainty is rather large for accurate diagnostic tests. We have to conclude that the usual method for testing the strength of association between the x-data and y-data in a linear regression model, although widely applied for validating quantitative diagnostic tests, seems to be inaccurate. Obviously, stricter criteria have to be applied for validation.

3. LINEAR REGRESSION TESTING THE HYPOTHESES THAT THE A-VALUE = 0.000 AND THE B-VALUE = 1.000

A stricter method to test the association between the new-test-data (the x-data) and the control-test-data (y-values) was given by Barnett.[5] First, from the above equation $y = a + bx$ it is tested whether the b-value is significantly different from zero like described above. Then, the hypothesis is tested that the a-value = 0.000 and the b-value = 1.000. As an example the graph from Figure 1 is used once more. We need the b-value (or a-value) ± 1.96 times its standard error to calculate the 95 % confidence intervals of b and a.

If the 95% confidence interval of the b-value (2.065 ± 1.96 x 0.276) contains 1.000,

and the a-value (8.647 ± 1.96 x 3.132) contains 0.000,

=> then validity can be accepted.

Here: the 95% confidence interval of the b-value is between 1.513 and 2.617,

and of the a-value is between 2.383 and 14.911,

=> the test can not be validated.

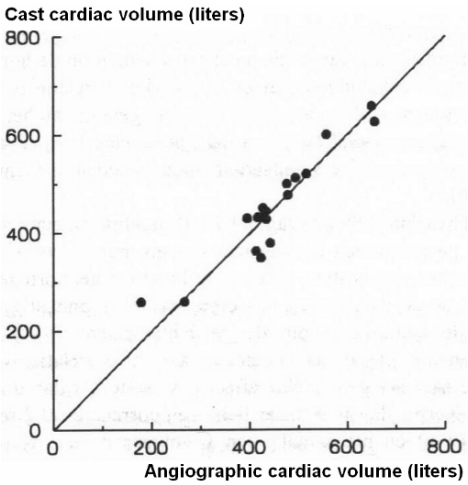


Figure 2. Angiographic cardiac volumes (liters) used to predict cast cardiac volumes (liters).

In the example of Figure 2 the data are close to the "b =1.000 and a = 0.000 line", otherwise called the identity line.

The 95% confidence interval of the b-value is between 0.917 ± 1.96 x 0.083,
is between 0.751 and 1.083,
and it, thus, contains the number 1.000.

The 95% confidence interval of the a-value is between 39.340 ± 1.96 x 38.704,
is between -38.068 and 116.748,
and it, thus, contains the number 0.000.

The diagnostic test of Figure 2 is validated. If the hypothesis that a = 0.000 and b = 1.000 can not be confirmed, and the b-value is significantly larger than 0, then the underneath method can be applied for validation. A b-value significantly smaller than 1.000 is an indicator for a diagnostic test that systematically overestimates the gold standard test, and significantly larger than 1.000 it is so for a diagnostic test that systematically underestimates the gold standard test.

4. LINEAR REGRESSION USING A SQUARED CORRELATION COEFFICIENT (R^2 – VALUE) OF > 0.95

The previous method assumes that the best fit linear regression equation for the diagnostic test is y = x. A diagnostic test with the best fit equation y = a + b x, rather than y = x like in the example from Figure 1 is not necessarily useless, and

could be approved as a valid test if it is precise, that means if the x-data precisely predict the *(y-a) / b- data* rather than the y-data. If we apply such a test, the result of the x-data will, of course, have to be transformed into $a + b\,x$ to find the y-data. Validation is accomplished by determining whether the regression line precisely predicts the control test from the new test, and we recommend to use for that purpose as criterion a squared correlation-coefficient $r^2 > 95\%$. This can be calculated from

$$r^2 = SP^2\,x\,y \;/\; (SS\,x \,.\, SS\,y).$$

In the example from Figure 1 $r^2 = 63\ \%$, much smaller than 95%, and, so, the results can not be validated. In the literature often the term *intraclass correlation* is applied instead of the r^2 – value, but its meaning is the same.

$$\text{Intraclass correlation} = \frac{\text{SS regression}}{\text{SS regression} + \text{SS residual}} = \frac{SP^2\,xy / SS\,x}{SS\,y}.$$

A largely similar approach is given by the calculation of the *relative residual variance* of the linear regression. The relative residual variance is calculated from the add-up sum of the least squared distances from the regression line (Figure 3)

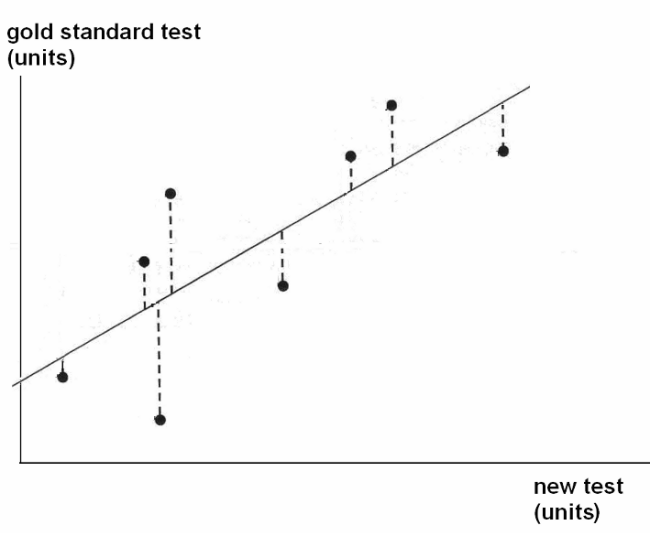


Figure 3. The relative residual variance is calculated from the add-up sum of the least squared distances from the points to the regression line.

and is equal to $(1 - r^2)$. The larger it is, the poorer the validity of the test. A residual variance smaller than 5% is adequate for validation.

$$\text{Relative residual variance} = \frac{\text{SS residual}}{\text{SS regression} + \text{SS residual}}$$

$$\text{SS residual} = SSy - (SP^2\,xy / SSx)$$

$$\text{Relative residual variance} = \frac{SSy - (SP^2\,xy / SSx)}{SSy} = \frac{SSy}{SSy} - \frac{(SP^2\,xy / SSx)}{SSy} = (1 - r^2)$$

= 37% in the example from Figure 1.

The levels 95% and 5 % are, of course, arbitrarily chosen. However, they are consistent with the cut-off for type I errors as commonly chosen in clinical research.

When using the linear model for testing validity, it is recommended to test the linear hypothesis, i.e. to test that the relationship between the new-test-data and control-test-data are, indeed, linear, rather than curvilinear. This can be done by testing the hypothesis that a second order correlation exists between the x- and y-data. For that purpose the equation $y = a + b\,x^2$ is used. If the data are modeled according to this equation and the b-value is significantly larger than the b-value of the linear model $y = a + bx$, then the linear model has to be rejected, and the data should be modelled according to a second order relationship between the new-test-values (x-values) and the control-test-values (y-values). If a log linear relationship between the diagnostic test and the gold standard test better fits the data than a linear relationship, then a so-called pseudo-R2 or R2-like measure instead of r^2 value can be calculated.[6]

5. ALTERNATIVE METHODS

All of the methods discussed so far assume uncertainty in the new test, but not in the control test. Two assessments, that assume uncertainty of both the new-test- and the control-test-data, are the paired Student's t-test and the Altman-Bland plot or method. The first uses the average difference between the new-test-values and the old-test-values as estimate of bias, and the standard error of the mean difference as estimate of precision.[7] The second [8] uses the spread of the subtraction sums of the new-test- and old-test-data and their standard deviation. If 95% of the subtraction sums fall within the limits of agreements, as calculated by the mean differences ± 1.96 times its standard deviation, then the test is validated. It may, generally, be perfectly all right to assume no uncertainty in the control test, particularly, if it is the gold standard test, for which there is no better alternative. The gold standard test, then, simply produces the truth. In this situation the additional amount of uncertainty assumed in the control test causes loss of sensitivity of testing. Even if the control test is not 100 % accurate, we are, generally, merely interested in the validation against the control test, no matter its accuracy. A second problem with the above two methods is that, unlike linear regression, they assume Gaussian-like sampling distributions of the subsequent x-

and y-values. This assumption is not always appropriate, since the data are, generally, not randomly sampled, but obtained from selected groups.

If we want to account the uncertainty of a control test, which is not a gold standard test, then a better approach will be to test both the new and the control test against the gold standard test. This will unmask which of the two tests performs better. In the situation where there is no certain gold standard test and where it is decided to account uncertainty of the control test to be used, Deming[9] and Passing-Bablok[10] regression are sometimes used instead. They are methods based on linear regression and mathematically more complex than simple linear regression. Deming regression, just like the paired t-test and the Altman-Bland plot, assumes normal distributions of the subsequent x- and y-data. In contrast, Passing-Bablok regression does not. It is a non-parametric method using the Kendall's rank-correlation test to assess the above-described hypotheses that b = 1.000, and a = 0.000. First, one should produce a ranked sequence of all possible slope-values between two x and two y-values (Sij values). We, then, compare the Sij values > 1 with those <1, and test whether there is a significant difference using Kendall's standard error equation, SE (standard error) = $\sqrt{\ }$ n(n-1) (2n+5) /18 with n = number of paired values. If, after continuity correction (add -1 to the difference as calculated), the SE is smaller than half the size of the calculated difference, then the b-value is not significantly different from 1.000. The a-value is calculated from the medians of x and y using the calculated b, its SE from the upper and lower limit of the confidence intervals of the b-value. The method is laborious, particularly, with large samples, but available through S-plus, Analyse-it, EP Evaluator, and MedCalc and other software programs.

6. DISCUSSION

Simple linear regression testing the presence of a significant correlation between the new-test-data (x-axis-data) and the control-test-data (y-axis-data) is not accurate for testing the validity of a novel quantitative diagnostic test. Accurate methods using linear regression include the following.

1. From y = a + b x, test the hypothesis that b is statistically significantly larger than zero, than test the hypothesis that b = 1.000 and a = 0.000.
2. If "the b = 1.000 and a = 0.000 hypothesis" cannot be confirmed, then use as criterion for validation a squared correlation-coefficient r^2 or intraclass correlation of > 95%, or a relative residual variance of < 5%. If the new test is validated this way, then the predicted control-test-values are calculated from the equation y = a + bx.

Altman-Bland plots, paired t-tests, Deming regression and Passing-Bablok regression assume uncertainty of both the new test and the control test. This is rarely a condition for validation, and carries the risk of unneeded loss of sensitivity of testing. However, if there is no gold standard test and it is decided to account the uncertainty of the control test, then Passing-Bablok regression is the only method adequate for non-normal data as often present in practice.

When using a data plot with one test on the x- and one on the y-axis, sometimes non-linear or curvilinear or exponential patterns can occur. The diagnostic test may, then, be useful even so. But, we will first have to find the best fit equation for the data, which is generally the equation producing the largest regression coefficient, and may, for example, look like $y = \log x$, $y = a + b\,x^2$ and many more forms. Such a test can be approved as valid, if it is a precise predictor of the control test, even if the x-data do not predict y, but rather something like antilog y or $\sqrt{}$ [(y-a)/b]. In practice, however, linear relationships are the most common pattern observed with quantitative diagnostic tests.

7. CONCLUSIONS

Clinical research is impossible without valid diagnostic tests. The methods for validating *quantitative* diagnostic tests have not been agreed upon by the scientific community. This chapter reviews the advantages and disadvantages of methods that could be used for that purpose. Using real data examples we review seven possible methods.*:*

Simple linear regression testing the presence of a significant correlation between the new-test-data (x-axis-data) and the control-test-data (y-axis-data) is not accurate for testing the validity of a novel quantitative diagnostic test. Accurate methods using linear regression include the following. First, from $y = a + b\,x$, test the hypothesis that b is statistically significantly larger than zero, than test the hypothesis that $b = 1.000$ and $a = 0.000$. Second, if "the $b = 1.000$ and $a = 0.000$ hypothesis" cannot be confirmed, then use as criterion for validation a squared correlation-coefficient r^2 or intraclass correlation of > 95%, or a relative residual variance of < 5%. If the new test is validated this way, then the predicted control-test-values are calculated from the equation $y = a + bx$.

The above three methods assume uncertainty of the new-test-data, but not of the control-test-data. Deming regression, Passing-Bablok regression, paired Student's t-tests, and Altman-Bland plots assume uncertainty of both the new test and the control test. This is rarely a condition for validation, and carries the risk of unneeded loss of sensitivity of testing. However, if the control test is not the gold standard test and it is decided to account the uncertainty of the control test, then Passing-Bablok regression is the only method that adjusts for non-normal data as frequently observed in practice.

More information on accuracy assessments of quantitative diagnostic tests is given by the CLSI protocols published by the Clinical and Laboratory Standards Institute, particularly the protocols EP9 and EP14.[11]

8. REFERENCES

1.Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995; 274: 645 -50.

2. Anonymous. The quality of diagnostic tests statement, the CONSORT

statement. www.consort-statement.org
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig JG, Moher D, Rennie D, De Vet HC, for the STARD steering group. Education and Debate. Towards Complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326: 41-4.
4. Delong ER, Delong DM. Comparing the areas under two or more correlated receiver operated characteristic curves; a nonparametric approach. Biometrics 1988; 44: 837-45.
5. Barnett DV. Simultaneous pairwise linear structural relationships. Biometrics 1969; 28: 129-42.
6. Hoetker G. The use of logit and probit models in strategic management research. Strat Mgmt J 2007; 28: 331-43.
7. McGee WT, Horswell JL, Calderon J, Janvier G, Van Severen T, Van den Berghe G, Kozikowski L. Validation of a continuous arterial pressure-based cardiac output measurement: a multicenter, prospective trial. Critical Care 2007; 11: R 105; pp 1-7.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i: 307-10.
9. Linnet K. Performance of Deming regression analysis in case of miss-specified analytical error in method comparison studies. Clin Chem 1998; 44: 1024-31.
10. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. J Clin Chem Clin Biochem 1983; 21: 709-20.
11. Guidelines for global application developed by the Clinical and Laboratory Standards Institute. www.techstreet.com.

CHAPTER 37

SUMMARY OF VALIDATION PROCEDURES FOR DIAGNOSTIC TESTS

1. INTRODUCTION

Clinical studies are impossible without adequate diagnostic tests, and diagnostic tests can, therefore, be considered the real basis of evidence-based medicine. In 1995 Reid et al[1] stated after a search of 1302 diagnostic studies that most diagnostic tests are inadequately appraised. Efforts to improve the quality of diagnostic tests are given by initiatives like those of the CONSORT[2] (Consolidated Standard Randomized Trials) movement and the STARD[3] (Standards for Reporting Diagnostic Accuracy) group launching quality criteria statements for diagnostic tests in 2002 and 2003. In spite of such initiatives the evaluation of diagnostic tests prior to implementation in research programs, continues to be lacking.[4] A diagnostic test can be either qualitative, e.g., the presence of an elevated erythrocyte sedimentation rate to demonstrate pneumonia, or quantitative, e.g., the ultrasound flow velocity to estimate the invasive electromagnetic flow velocity. For both qualitative and quantitative diagnostic tests three determinants of validity have been recommended by working parties:
Assess accuracy: the test shows who has the disease and how severe it is.
Assess reproducibility: when a subject is tested twice, the second test produces the same result as the first test.
Assess precision: there is a small spread in a random sample of test results.
The methods of assessment have, however, not been defined so far. The current paper reviews correct and incorrect methods and new developments.

2. QUALITATIVE DIAGNOSTIC TESTS

Accuracy

Assessing accuracy is probably most important. Accuracy is synonymous to validity, and can here be defined as a test's ability to show which individuals have the disease and which do not. It is, generally, assessed by sensitivity and specificity, defined as the chance of a true positive and true negative test respectively.
How do we calculate accuracy

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 433–447.

Disease	yes (n)	no (n)
Positive test	a	b
Negative test	c	d

n = number of patient

a = number of true positive patients
b = false positive patients
c = false negative patients
d = true negative patients

Sensitivity of the above test = a / (a+c)
Specificity = d / (b+d)
In addition to sensitivity and specificity sometimes overall accuracy is given
Overall accuracy = (a+d) / (a+b+c+d)

It is important to realize that a sensitivity/specificity close to 50% gives no more information than does flipping a coin, and that such a result is not a basis for validation. Often qualitative diagnostic tests have multiple sensitivities / specificities dependent on normal values used. In the example of the Figures 3 and 4 in Chapter 33 the erythrocyte sedimentation rate (ESR) is used as an estimator of pneumonia with chest x-ray as gold standard test. The sample population consists of two Gaussian distributions of patients, one with and the other without pneumonia.
Figure 1 in Chapter 33 shows that, if a normal value of the ESR is defined as < 43 mm, many healthy subjects are rightly diagnosed. However, many diseased are missed. The test, thus, produces a high specificity, but low sensitivity: we have many false negatives. If, in contrast, an ESR of >32 mm is used as level between health and disease (Figure 2, Chapter 33), then we do not miss many diseased, but we will misdiagnose many healthy subjects. Our test will have a low specificity: we will have many false positives. The question is what normal ESR value is best in order to miss as few diagnoses as possible, and obtain both a high sensitivity and high specificity. ROC (receiver operating) curves are helpful for finding both (Figure 1). First, we calculate for several tentative normal values sensitivity /

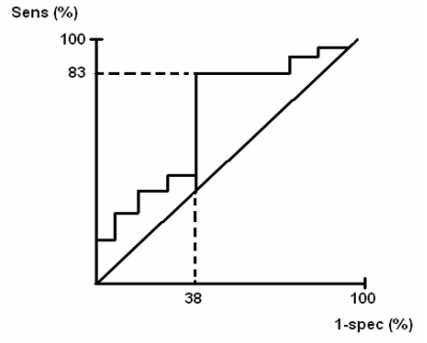


Figure 1. The ROC (receiver operating characteristic) curve of the erythrocyte sedimentation rate (ESR) values of the patients with pneumonia from the Figures 1 and 2 plot the sensitivity values against the "1-specificity" values (sens = sensitivity, spec = specificity).

specificity. Then, we draw a curve with sensitivities on the y-axis and specificities or 1-specificities (producing a somewhat prettier curve) on the x-axis. A perfect test reaches the top of the y-axis where both sensitivity and 1-specificity are 100%. The given example does not produce a perfect test, but we can readily observe from the graph that an ESR of 38 mm produces the shortest distance to the top of the y-axis. If you want proof, you may wish to measure the distance between the top of the y-axis and the curve or calculate it using rectangular triangles and Pythagoras' equation.

ROC curves are very popular, but have some limitations. First, sometimes more than 1 shortest distance to the top of the y-axis is observed. Second, ROC curves close to the diagonal provide no more information than tossing a coin (overall accuracy is only 50%). Third, often two different diagnostic tests are compared for identifying the better of the two using areas under the curve of the ROC curves. A problem is, that such ROC curves often cross, which means, that a diagnostic test may perform better in one interval, worse in another.

Reproducibility

Cohen's kappas are used for assessing reproducibility of qualitative diagnostic tests. As an example 30 patients are assessed twice for a positive test for brain natriuretic peptide for a diagnosis of heart failure.

		1st time positive test		
		yes	no	
2nd time positive test	yes	10	5	15
	no	4	11	15
		14	16	30

It can be demonstrated that, if the test were not reproducible at all,

you would find (14 x 15 / 30 =) 7 x twice yes
and (16 x 15 / 30 =) 8 x twice no +
15 x twice the same.

In fact, we do find 21 x twice the same.

The kappa estimator is calculated according to:

$$\text{Kappa} = \frac{\text{observed - minimal}}{\text{maximal - minimal}} = \frac{21 - 15}{30 - 15} = 0.4$$

The result is interpreted as follows: a kappa value of 0 means a very poor reproducibility, a value of 1 an excellent reproducibility. In our example the reproducibility is moderate.

Precision

The STARD (standards for reporting diagnostic accuracy) working party[5] has proposed to include a "measure of uncertainty" in any validation procedure. Standard deviations / errors (SDs / SEs) can be used for that purpose.

Disease	yes (n)	no (n)
Positive test	a	b
Negative test	c	d

$SE_{sensitivity} = \sqrt{ac/(a+c)^3}$
$SE_{specificity} = \sqrt{db/(d+b)^3}$
$SE_{overall\ accuracy} = \sqrt{prev^2 \times var_{sensitivity} + (1-prev)^2 \times var_{1-specificity} + (sens-spec)^2 \times var_{prev}}$
where prev = prevalence = (a+d) / (a+b+c+d) and var = variance = SD^2.

A small sensitivity with a relatively wide spread, for example, a sensitivity of 55% with a $SE_{sensitivity}$ larger than 2.5% means that the sensitivity is not significantly different from 50%, a result that does not give more information than tossing a coin. This diagnostic test is not adequately precise for validation. Many diagnostic tests have been erroneously validated in the past based on sensitivities / specificities higher than 50%, without assessment for uncertainty (see Chpater 34).

3. QUANTITATIVE DIAGNOSTIC TESTS

Accuracy

Linear regression with the gold standard test as dependent and the new diagnostic test as independent variable (respectively y- and x-variable) is very popular for assessing accuracy of quantitative diagnostic tests. If a statistically significant association between y and x is established, this is generally considered sufficient evidence for validation. This approach is incorrect. For example, in Figure 2 an

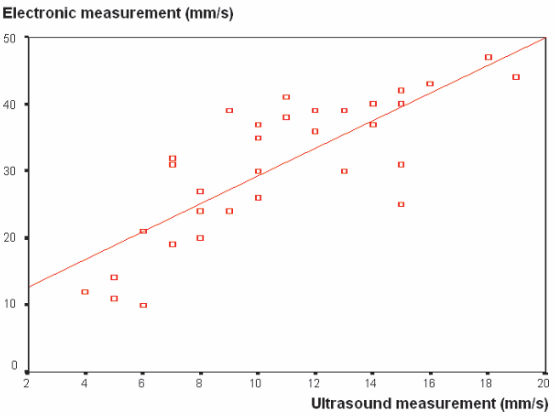


Figure 2. Accuracy assessment with a linear regression model. The regression equation is given by y = a + b x = 8.647 + 2.065 x (a= intercept, b = regression coefficient, p <0.0001). The x-variable, ultrasound estimate, is a very significant predictor of the y-variable, the electromagnetic measurement. However, the prediction, despite the high level of statistical significance, is very imprecise. E.g., if x = 6, then y may be 10 or 21, if x = 7, y may be 19, 31 or 32.

example is given where flow velocity (mm / s) as estimated by ultrasound is used to predict the standard electromagnetic measurements (mm / s). The regression equation calculated from the SPSS Statistical Software program is given by y = a + b x = 8.647 + 2.065 x (a= intercept, b = regression coefficient). The standard error of the regression coefficient b = Se_b = 0.276. This means that the t-value = 2.065 / 0.276, equaling 7.491, and the p-value is thus <0.0001. The x-variable, ultrasound

estimate, is a very significant predictor of the y-variable, the electromagnetic measurement. However, the graph shows that the prediction despite the high level of statistical significance, is very imprecise. E.g., if x = 6, then y may be 10 or 21, if x = 7, y may be 19, 31 or 32. A significant correlation is, thus, not good enough to validate a quantitative diagnostic test. A more adequate method for validation was given by Barnett.[5] Test the hypotheses a = 0.000, and b = 1.000. If the 95% confidence intervals of the calculated a and b include the numbers 0.000 and 1.000 respectively, then the test can be accepted as validated. Confidence intervals can be calculated several ways, but here we use the Gaussian approach:
95% confidence interval of a = a ± 2 Se_a (8.647 ± 2 x 3.132)
95% confidence interval of b = b ± 2 Se_b (2.065 ± 2 x 0.276)
a ± 2 Se_a = between 2.383 and 14.911.
b ± 2 Se_b = between 1.513 and 2.617
The numbers 0.000 and 1.000 are not included in the 95% confidence intervals. No validity has been established.

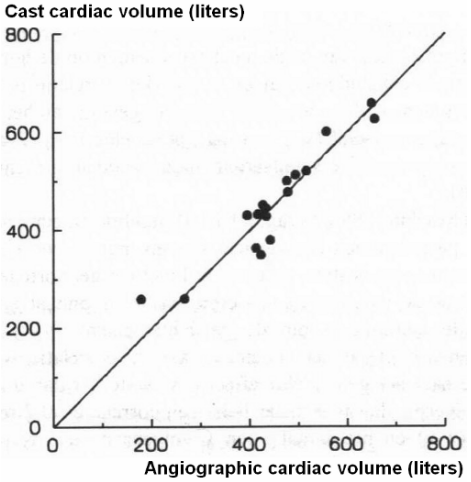


Figure 3. Angiographic cardiac volumes (liters) used to predict cast cardiac volumes (liters).

Another example is given in Figure 3. Angiographic cardiac volumes (liters) are used to predict cast cardiac volumes (liters). When testing the hypotheses a = 0.000 and b = 1.000, SPSS will produce the following results.
95 % confidence intervals of a = a ± 2 Se_a =39.340 ± 2 x 38.704
95% confidence intervals of b = b ± 2 Se_b = 0.917 ± 2 x 0.083
a ± 2 Se_a = between -38.068 and 116.748
b ± 2 Se_b = between 0.751 and 1.083

The 95% confidence intervals include 0.000 and 1.000 respectively. The diagnostic can be accepted as validated.

Reproducibility

Reproducibility is often calculated incorrectly. The first commonly used incorrect method is given in the example underneath. The individual differences between test 1 and 2 per patient are calculated. If the mean difference is small, it is concluded that the test is well reproducible.

Patient no	test 1	test 2	difference
1	1	11	- 10
2	10	0	10
3	2	11	-9
4	12	2	10
5	11	1	10
6	1	12	-11
		Mean difference	0

As can be observed from the above example of flow velocities (mm/s), the mean difference between test 1 and test 2 is zero. Yet, the tests are very poorly reproducible, because the range of differences is no less than 21 (differences vary from -11 to +10) mm/s.
The second commonly used incorrect method is the following. A regression line is drawn with test 1 data on x-axis and test data on y-axis. If the data are close to the line, it is concluded that reproducibility is good. There are two problems with this approach. First, testing twice introduces a regression to the mean phenomenon: patients scoring low the first time, have a better chance of scoring higher next time vice versa. A second problem is that only good reproducibility is an adequate conclusion if the direction coefficient of the regression line has a direction of 45 degrees.
The correct methods for assessing reproducibility with quantitative diagnostic tests are summarized.
1.Duplicate standard deviation
2.Repeatability coefficient
3.Intraclass correlation

1.Duplicate standard deviation (SD)
The duplicate standard deviation is used in the underneath example.

Patient no	test 1	test 2	difference(d)	$(\text{difference})^2$
1	1	11	- 10	100
2	10	0	10	100
3	2	11	- 9	81
4	12	2	10	100
5	11	1	10	100
6	1	12	-11	121
average	6.17	6.17	0	100.3

Duplicate SD = $\sqrt{\ }$ ½ Σ d^2 /n = $\sqrt{\ }$ (1/2 x 100.3) = 7.08
Duplicate SD % = $\frac{\text{duplicate SD}}{\text{overall mean}}$ x 100% = $\frac{7.08}{6.17}$ x 100% = 115%
An adequate reproducibility corresponds to a duplicate SD of 10-20%.

2.Repeatability coefficient
The repeatability coefficient is applied in the underneath example.

Patient no	test 1	test 2	difference
1	1	11	- 10
2	10	0	10
3	2	11	- 9
4	12	2	10
5	11	1	10
6	1	12	-11
Mean	6.17	6.17	0
Standard deviation (SD)			10.97

The repeatability coefficient = mean difference ± 2 $SD_{\text{difference}}$ = 0 ± 21.95
The interpretation is as follows. A repeatability coefficient must be < than the largest measured difference between test 1 and test 2.

3.Intraclass correlation
The intraclass correlation is applied in this example (SD = standard deviation, SS = sum of squares)).

patient	test 1	test 2	average	SD^2
1	1	11	6	50
2	10	0	5	50

3	2	11	6.5	40.5
4	12	2	7	32
5	11	1	6	50
6	1	12	6.5	60.5
mean	6.17	6.17		
overall mean	6.17			

$SS_{\text{between subjects}}$ = (mean subject 1 – grand mean)^2 + (mean subject 2 – grand mean)^2 + …….. = 3.0134
$SS_{\text{within subjects}} = SD_1^2 + SD_2^2 + SD_3^2 + SD_4^2 + \ldots = 283$

The intraclass correlation (ICC) is given be the equation =

$$\frac{SS_{\text{between subjects}}}{SS_{\text{between subjects}} + SS_{\text{within subjects}}} = 0 - 1$$

If $SS_{between} = 0$, then the test will be poorly reproducible; if $SS_{within} = 0$, then the test will be excellently reproducible. Here the intraclass correlation = 0.01051, and, so, the test is very poorly reproducible.

Precision

A good precision can be interpreted as a small spread in the data, for example, estimated by a small SD or SE (standard error). If the spread in a data sample is wide, some legitimate statistical methods are available to reduce the size of the SDs / SEs, such as data modeling (massage) using multiple regression or logarithmic transformation, exponential modeling, polynomial modeling or other methods. Figure 4 gives an example of data modeled using a multiple linear

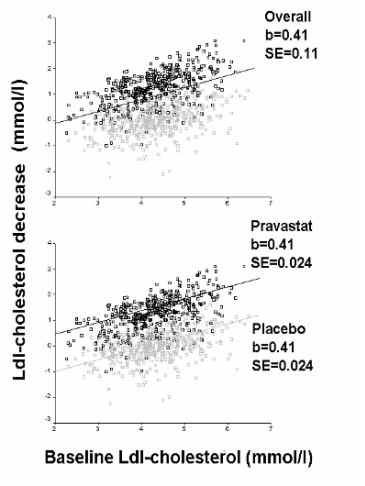


Figure 4. Example of data modeling for increasing precision using a multiple linear regression model. On the x-axis the baseline Ldl-cholesterol level of the patients in a cholesterol-study are given, on the y-axis the decrease of Ldl-cholesterol after treatment is given. The upper graph gives the results without, the lower with modeling. It can be observed in the figure that the treatment efficacy given by the b-values are similar but that the SEs are smaller in the modeled graph, and so this modeling produced a better precision.

regression model. On the x-axis the baseline Ldl-cholesterol levels of the patients in a cholesterol-study are given, on the y-axis the decreases of Ldl-cholesterol after treatment is given. The upper graph gives the results without, the lower with modeling. It can be observed in the figure that the treatment efficacy given by the b-values is unchanged, but that the spread in the data given by the SEs is smaller with the multiple linear regression models.

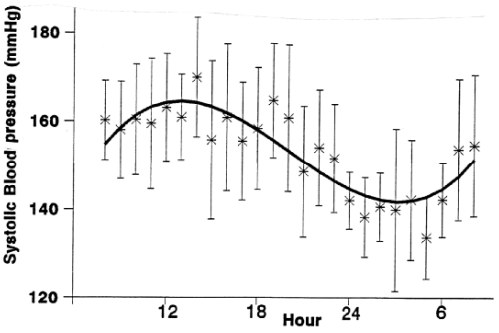


Figure 5. Ambulatory blood pressure measurements (ABPMs) of 10 subjects; both means and SDs and 7th order polynomial regression models of the data are drawn. The SPSS program calculates the spread in the data using either of the two methods. The pooled SD of the ABPM values using the means equals 17 mm Hg (pooled departure from all means). The SD of the polynomial model is much smaller and equals 7 mm Hg (pooled departure from the polynomial curve). Obviously, this curvilinear regression model provides a much better precision, and is, therefore, a more precise model for analyzing the differences between the ABPM recordings of different blood pressure reducing therapies than simply the use of averages and their SDs.

Another example is in Figure 5 showing the ambulatory blood pressure measurements (ABPMs) of 10 subjects; both means and SDs and 7th order polynomial regression models of the data are drawn. The SPSS program calculates the spread in the data using either of the two methods. The Pooled SD of the means equals 17 mm Hg. The SD of the polynomial model is much smaller and equals 7 mm Hg. Obviously, this curvilinear regression model provides a better precision, and is, therefore, a more precise model for analyzing the effects on the ABPM recordings of different blood pressure reducing therapies than simply using the averages and their SDs.

4. ADDITIONAL METHODS

Three relatively new methods for the accuracy assessment of diagnostic tests are available.

1.Continuous receiver operated characteristic (ROC) curves[6,7]

In the use of many diagnostic tests, test results do not necessarily fall into one of two categories, but rather into categories with more or less confidence in the presence of disease (Figure 6). While using multiple thresholds for making a

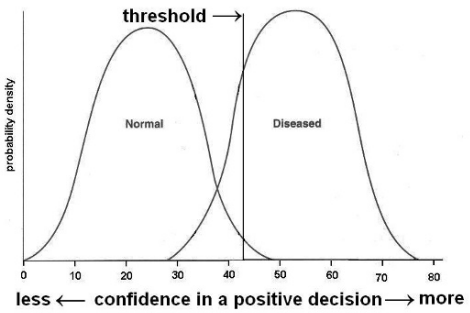


Figure 6. Diagnostic test where results do not fall into one of two categories but rather into categories with more or less confidence.

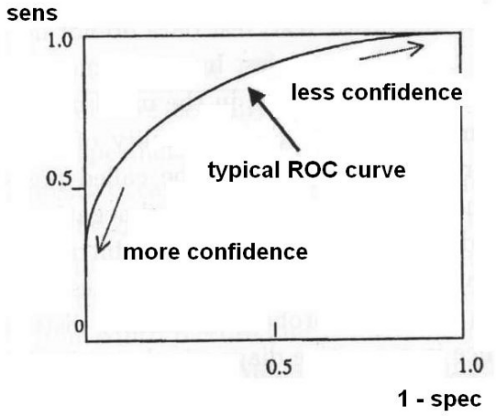


Figure 7. Continuous receiver operated characteristic (ROC) curve.

diagnosis, a continuous ROC curve can be obtained (Figure 7) The procedure is somewhat laborious, because appropriate software is not yet available, but the procedure has the advantage that a precise area under the curve can be calculated. The closer this ROC curve approaches the top of the y-axis, the better accuracy the diagnostic test will provide with an optimal AUC of 1.0. In contrast, if the ROC curve is close to the 45 degree diagonal line, the AUC is close to 0.5, and the test is very inaccurate. The simplest method for calculating the AUC is summing the areas of the trapezoids formed by the curve and the x-axis.

2.Intraclass correlation (ICC) for agreement with the gold standard test[8]

The Intraclass correlation has recently been used not only for assessing reproducibility of diagnostic tests, but also as an alternative method for accuracy assessments. It is given by the equation (SS = sum of squares)

$$ICC = \frac{SS_{\text{between techniques}}}{SS_{\text{between techniques}} + SS_{\text{within techniques.}}}$$

The interpretation is similar to that of the intraclass correlation method for reproducibility assessment as explained above. It can be demonstrated that this method is less sensitive in demonstrating disagreement between the diagnostic test and the gold standard test than the previously mentioned methods. Also the Bland-Altman method, which will be discussed next, performs better.

3.Bland-Altman method[9]

Bland and Altman recommended the following approach. Calculate the individual differences between the diagnostic test results and the gold standard test results and, subsequently, the standard deviation of these differences. If this standard deviation is less than or equal to the standard deviation of the both the diagnostic test results and the gold standard test results, then the two difference tests are exchangeable and, therefore, equivalent. The diagnostic test is, then, accurate.

5. DISCUSSION

The current paper gives some relatively simple methods for assessment. Validity assessments of diagnostic tests are rarely communicated in research papers and this may contribute to the low reproducibility of clinical trials. We expected that validation would, at least, be a standard procedure in clinical chemistry studies where a close to 100% accuracy / reproducibility is not unusual. However, even in a journal like the Journal of the International Federation of Clinical Chemistry and Laboratory Medicine out of 17 original papers publishing novel chemistry methods in 2006 none of the papers communicated validity assessments except for one study.[10] Ironically, this very study reported two incorrect methods for assessing reproducibility, namely the assessment of significant differences between repeated measures, and the calculation of Pearson's correlation levels.

A more general explanation for the underreporting of validation procedures for diagnostic tests in research communications is that the scientific community although devoted to the study of disease management, is little motivated to devote its energies to assessing the validity of the diagnostic procedures required for the very study of disease management. Clinical investigators favor the latter to the former. Also the former gives no clear-cut career path, while the latter more often does so. And there is the injections from the pharmaceutical industry. To counterbalance this is a challenge for governments and university staffs. Correct

methods for validation of both quantitative and qualitative diagnostic methods are summarized in Table 1.

Table 1. Summary of correct methods for validation of both quantitative and qualitative diagnostic tests.

Accuracy	Reproducibility	Precision
	Qualitative diagnostic test	
Sensitivity	Kappas	Confidence intervals
Specificity		
Overall accuracy		
ROC curves		
	Quantitative diagnostic test	
Barnett's test	Duplicate standard deviation	Confidence intervals
Intraclass correlation vs gold test	Repeatability coefficient	
Bland-Altman test	Intraclass correlation vs duplicate test	

6. CONCLUSIONS

Clinical developments of new treatments are impossible without adequate diagnostic tests. Several working parties including the Consolidated Standard Randomized Trials (CONSORT) movement and the Standard for Reporting Diagnostic Accuracy (STARD) group have launched quality criteria for diagnostic tests. Particularly, accuracy-, reproducibility- and precision-assessments have been recommended, but methods of assessment have not been defined so far.

This chapter summarizes correct and incorrect methods and new developments for that purpose.

A diagnostic test can be either qualitative like the presence of an elevated erythrocyte sedimentation rate to demonstrate pneumonia, or quantitative like ultrasound flow velocity to estimate invasive electromagnetic flow velocity.

Qualitative diagnostic tests can be assessed for:

-*accuracy* using sensitivity / specificity / overall accuracy, and receiver operated (ROC) curves, -*reproducibility* using Cohen's kappas, -*precision* using confidence intervals of sensitivity / specificity / overall accuracy.

Quantitative diagnostics tests can be assessed for

-*accuracy* using a linear regression line ($y = a + b\ x$) and testing $a = 0.00$ / $b = 1.00$, -*reproducibility* using duplicate standard errors, repeatability coefficients or

intraclass correlations, -*precision* by calculating confidence intervals. Improved confidence intervals can be obtained by data modeling.
A significant linear correlation between the diagnostic test and the gold standard test does not correctly indicate adequate accuracy. A small mean difference between repeated measures or a significant linear relationship between repeated measures does not indicate adequate reproducibility.
New developments include continuous ROC curves, intraclass correlations, and Bland-Altman agreement tests for the accuracy assessments of quantitative diagnostic tests.

7. REFERENCES

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995; 274: 645 -50.
2. Anonymous. The quality of diagnostic tests statement, the CONSORT statement. www.consort-statement.org
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig JG, Moher D, Rennie D, De Vet HC, for the STARD steering group. Education and Debate. Towards Complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326: 41-4.
4. Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large scale replication study. JAMA 2007; 297: 1551-61.
5. Barnett DV. Simultaneous pairwise linear structural relationships. Biometrics 1969; 28: 129-42..
6. Delong ER, Delong DM. Comparing the areas under two or more correlated receiver operated characteristic curves; a nonparametric approach. Biometrics 1988; 44: 837-45.
7. Hanley JA, McNeil BJ. The meaning and use of the area under curve under a receiver operating characteristic (ROC) curve. Diag Radiol 1982; 143: 29-36.
8. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989; 19: 61 -70.
9. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol 1995; 24: s7-s14.
10. Imburt-Bismut F, Messous D, Thibaut V, Myers RB, Piton A, Thabut D, Devers L. Hainque B, Mecardier A, Poynard T. Intra-laboratory analytical variability of biochemical markers of fibrosis and activity and reference ranges in healthy blood donors. Clin Chem Lab Med 2004; 42: 323-33.

CHAPTER 38

VALIDATING SURROGATE ENDPOINTS OF CLINICAL TRIALS

1. INTRODUCTION

Clinical trials are often constructed with surrogate endpoints for practical or cost considerations, for example, lipid levels as a surrogate for arteriosclerosis, arrhythmias for coronary artery disease, and cervical smears for tubal infections.[1-5] Such trials make inferences from surrogate observations about the effect of treatments on the supposed true endpoints without accounting the strength of association between the surrogate and true endpoints. The main problem with this practice is that the surrogate endpoint may lack sufficient validity to predict the true endpoint, giving rise to misleading trial results. The International Conference of Harmonisation (ICH) Guideline E9 Statistics Principles for Clinical Trials[6] recommends that, for the approval of a surrogate marker, (1) a statistical relationship with the true endpoint in observational studies be demonstrated, (2) evidence be given from clinical trials that treatment effects on the surrogate correspond to those on the true clinical endpoint, and (3) the surrogate marker like a diagnostic test be tested for sensitivity and specificity to predict the true endpoint. There is, thus, considerable consensus to routinely assess the accuracy of surrogate markers, but not specifically how to do so. Problems with the current sensitivity-specificity approach to validity is, that it is dual and that an overall level of validity is, therefore, hard to give.[7] Also, it can be used for binary (yes / no) endpoints only. As an alternative, regression-models have been proposed.[6,8] However, a correlation of borderline statistical significance between the surrogate and the true endpoint is not enough to indicate that the surrogate is an accurate predictor. The current paper underscores the need for accuracy assessment of surrogate endpoints by comparing the required sample sizes of trials with and without surrogate endpoints, and describes two novel procedures for assessment. The first makes use of an overall level of accuracy with confidence intervals and a prespecified boundary of accuracy. The second uses a regression model that accounts both the association between the surrogate and the true endpoint, and the association between either of these variables and the treatments to be tested.

2. SOME TERMINOLOGY

Surrogate marker/endpoint/test	Laboratory measurement or physical sign used as a substitute for a clinically meaningful endpoint that measure directly how a patient feels, functions, or survives, otherwise called the true endpoint.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 449–459.

Validity of a surrogate test	The surrogate test's ability to show which individuals have a true test either positive or negative. We sometimes use the term overall validity to emphasize that the approach is different from assessing sensitivity and specificity separately.
Sensitivity	Chance of a true positive surrogate test.
Specificity	Chance of a true negative surrogate test.
Odds ratio (OR)	Odds of the clinically meaningful endpoint in the treatment group / odds of it in the control group.
Alpha (α)	Type I error, chance of finding a difference where there is none.
Beta (β)	Type II error, chance of finding no difference where there is one.
Null-hypothesis	The study is negative, the treatment does not work. The null-hypothesis of no treatment effect is rejected when the difference from a zero effect is significant.
Variance	Estimate of spread or precision in the data. Variance of proportion p = p (1-p)
Standard error (SE)	$\sqrt{}$ (variance/n), where n = sample size.
Confidence interval (CI)	It covers a percentage of the results that can be expected if the study would be repeated many times. E.g., 95% CI between an OR of 1.10 and 1.86 means that 95% of many similar studies would produce an OR between 1.10 and 1.86. 95%CI of a proportion be calculated according to: proportion ± 1.96*$SE_{proportion}$, where * is the sign of multiplication.
Prespecified boundary of validity	It is often chosen on clinical grounds, and covers the range of results that are accepted by the investigators as sufficiently valid to use the surrogate test for its purpose. Currently, it is considered good statistical practice to define a prespecified boundary of your expected validity, and, then, test whether the confidence interval of your calculated level of validity falls entirely within the prespecified boundary. If so, you

	accept, if not you reject the presence of validity.
Dependent variable	y-variable in a regression analysis.
Independent variable	x-variable in a regression analysis.
Correlation coefficient squared (r^2)	Estimate of strength of association between paired observations. If $r^2 = 0$, there is no association, if $r^2 = 1$, there is 100% association. If $r^2 = 0.5$, there is 50% association. One variable determines the other by 50%, and there is 50% uncertainty. The r^2-value expresses the proportion of variability in the y-variable determined by the variability in the x-variable.
Regression coefficient (b)	Estimate of strength of association between paired observations particularly used in the case of multiple regression.

3. SURROGATE ENDPOINTS AND THE CALCULATION OF THE REQUIRED SAMPLE SIZE IN A TRIAL

The validity or accuracy of a surrogate marker can be expressed in terms of sensitivity and specificity to predict the true endpoint, e.g. healings.

	healings	non-healings
new treatment (group1)	170 (E)	140 (F)
control treatment (group 2)	190 (G)	230 (H)

odds of healing E/F and G/H,
odds ratio (OR) = E/F / G/H
= (170/140)/(190/230) = 1.47.

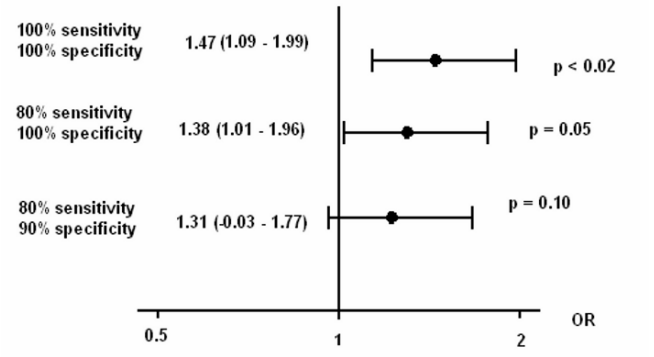


Figure 1. Effect of sensitivity and specificity levels on odds ratios and their 95% confidence intervals (odds ratio = odds of healing of the new treatment / odds of healing of the control treatment).

Figure 1 shows that a true endpoint test for the assessment of the above data has a 95% confidence interval between 1.09 and 1.99, and that it can reject the null-hypothesis of no difference between the two treatments at P<0.02. If a surrogate test for the assessment of the same data has a sensitivity of 80% and specificity of 100%, the OR will diminish, because the observed numbers of healings will fall by 20%, and those of the non-healings will rise correspondingly (OR = 1.38; 95% confidence interval 1.10-1.86, P = 0.05). If it has a sensitivity of 80% and specificity of only 90%, the OR can be calculated to further fall to 1.31 (95% confidence interval –0.029 to 1.77, p = 0.10), and a significance of difference between the two treatments can no longer be demonstrated (Fig.(**1**)). Obviously, with surrogate markers rapidly less certainty is provided to estimate the chance of healing or no-healing. In order to maintain a close to true endpoint level of certainty the sample size will have to be increased.
The effect on sample size requirement of a reduced sensitivity or specificity is illustrated in the underneath hypothesized example.

In a parallel study group 1 10% healings are expected,
group 2 20% healings are expected.

The required sample size can be calculated according to:

$$\text{required sample size} = \text{power index} * \frac{p_1(1-p_1)+p_2(1-p_2)}{(p_1-p_2)^2} \text{ subjects per group}$$

= 195 subjects per group

p_1 = expected proportion of healings in group 1, p_2 = expected proportion of healings in group 2, power index for $\alpha = 0.05$ and $\beta = 0.20$ equals 7.8, * = the sign of multiplication.

If the surrogate test provides 80% sensitivity, then in group 1 not 10% but 80% x 10% = 8% healings will be observed, in group 2 not 20% but 80 % x 20% = 16 %. The required sample size will rise to:

= 254 subjects per group.

If sensitivity = 80% and specificity = 90%, it can be similarly calculated that the required sample size will further rise to no less than:

= 515 subjects per group.

In trials using surrogate endpoints the sample size has to be based not only on the expected treatment efficacy but also on the validity of the surrogate marker used. We will now describe two procedures that can be readily applied for validating the surrogate marker. The first is adequate for binary variables, the second both for continuous and binary variables. The first can also be chosen after the assignment of continuous data to binary ones.

4. VALIDATING SURROGATE MARKERS USING 95% CONFIDENCE INTERVALS

The validity of a surrogate marker can, like a diagnostic test, be assessed by sensitivity and specificity to predict the true endpoint. In addition to this dual approach to accuracy, an overall validity can be calculated as illustrated below.

		observed surrogate endpoint (n)	
		yes	no
observed true endpoint	yes	a	b
	no	c	d

Sensitivity = a / (a+c)
Specificity = d / (b+d)
1-specificity = b / (b+d)
Prevalence of true endpoint = (a+b) / (a+b+c+d)
The variance of sensitivity is given by $ac / (a+c)^3$.
For the specificity the variance = $db / (d+b)^3$.
Also for 1-specificity the variance = $db / (d+b)^3$
For the prevalence of the true endpoint the variance = $(a+b)(c+d) / (a+b+c+d)^3$

Overall validity = sensitivity * prevalence + specificity * (1-prevalence).
* = the sign of multiplication.

For approval of a surrogate marker a boundary of validity is prespecified in the study protocol, e.g., 85% < validity < 100%, and a confidence interval of the overall validity level is calculated. If the confidence interval falls entirely between the prespecified boundary, validity is demonstrated. E.g., the true endpoint is a cardiovascular event, the surrogate endpoint is an elevated C-reactive protein level, currently a widely used marker for cardiovascular disease.
For the calculation of the confidence intervals standard errors (SEs) are required. In order to calculate the standard error (SE) (= √ (variance)) of the overall validity, we make use of the formula:

Var(X+Y) = Var(X) + Var(Y) + 2 Cov(X,Y).

Var (overall validity) =

Var(sens*prev) + Var(spec)*(1-prev) + 2Cov(sens*prev, spec*(1-prev)).

Var = variance; sens = sensitivity; spec = specificity; prev = prevalence; cov = covariance.

The variance of X*Y may be approached from

$Var(X*Y) = Y^2 Var(X) + X^2 Var(Y)$.

Using this formula we will end up finding:

Var(overall validity) =

$prev^2 * Var(sens) + (1-prev)^2 * Var(1-spec) + (sens-spec)^2 * Var(prev)$.

If, e.g.,

sensitivity = 80 % with SE = 2%,
specificity = 90% with SE =1%,
prevalence =10% with SE =3%,

then we can calculate:

overall validity = 0.8*0.1 + (0.9)*(1-0.1) = 0.89

and

$Var(\text{overall validity}) = 0.1^2 * 0.02^2 + (1-0.1)^2 * 0.02^2 + (0.8-0.9)^2 * 0.03^2$
$= 0.000337.$

The SE of the overall validity is the square root of the variance, and equals 0.018356 = 1.8356 %.

This approach makes use of the so-called delta-method which describes the variance of natural logarithm (ln) (X) as $Var(\ln(x)) = Var(x) / x^2$. The approach is sufficiently accurate if the standard errors of prev, sens and spec are small which is true if samples are large.

An overall validity of 89% with SE 1.8356 % means that the 95% confidence interval is between 0.89-1.96*0.018356 and 0.89 + 1.96*0.018356, and is thus between 85.4 and 92.6 %. This interval falls entirely between the prespecified interval of validity of 85% < validity < 100%. This surrogate endpoint is, thus, validated.

5. VALIDATING SURROGATE ENDPOINTS USING REGRESSION MODELING

Table 1. Total cholesterol and LDL-cholesterol levels are used as tentative surrogate endpoints for coronary artery diameter

Pt no.	Cor art (mm)	Treat	Tchol (mmol/l)	LDLchol (mmol/l)	Pt no.	Cor art (mm)	Treat	Tchol (mmol/l)	LDLchol (mmol/l)
1	24	0	4.0	2.4	18	12	0	2.0	1.0
2	30	0	6.5	3.2	19	26	0	5.0	2.8
3	25	0	7.5	2.4	20	20	1	4.0	2.0
4	35	1	5.0	3.6	21	43	0	8.0	4.4
5	39	1	4.5	3.8	22	31	0	7.5	3.0
6	30	0	5.0	3.0	23	40	1	7.0	3.8
7	27	0	4.0	2.6	24	31	0	3.5	3.2
8	14	0	2.5	1.6	25	36	1	6.0	3.4
9	39	1	6.5	4.0	26	21	0	3.0	2.0
10	42	1	7.5	4.2	27	44	0	9.5	4.6
11	41	1	5.5	4.0	28	11	1	2.5	1.0
12	38	1	5.5	3.8	29	27	0	4.0	2.6
13	39	1	6.0	3.6	30	24	0	4.5	2.6
14	37	1	5.0	3.4	31	40	1	7.5	3.8
15	47	1	9.0	4.8	32	32	1	3.5	3.4
16	30	0	6.5	2.8	33	10	0	3.0	0.8
17	36	1	6.0	3.8	34	37	1	7.0	3.2
					35	19	0	3.5	2.0

Pt no.= patient number; Cor art = coronary artery diameter; Treat = treatment modality (0 = placebo. 1 = active treatment); Tchol= total cholesterol level; LDLchol = LDL-cholesterol level

Table 1 shows the total and LDL cholesterol levels being used as tentative surrogate endpoints for coronary artery diameter. For the validation of the two surrogate endpoints the following linear model is used:

$$y = a + b_1 x_1 + b_2 x_2$$

y = true endpoint,
x_1 = treatment modality (0 = placebo, active treatment = 1)
x_2 = surrogate endpoint

$$y = a + b_1 x_1$$

r^2 of this equation = proportion variability in y explained by x_1

$y = a + b_1 x_1 + b_2 x_2$
r^2 of this equation = proportion variability in y explained by x_1 and x_2.

The subtraction sum of the two r^2-values = proportion variability y explained by the surrogate endpoint x_2; the larger the subtraction sum the better the surrogate endpoint. Table 2 gives a summary of the calculations. Both LDL-cholesterol and

Table 2. Analysis of associations between true endpoint, treatment modality and surrogate endpoints from Figure 1

	r^2 -value	F-value	p-value
True vs treat	0.250	10.9	0.002
LDL-chol vs treat	0.202	8.3	0.007
Tchol vs treat	0.044	1.5	0.226
True vs LDL-chol	0.970	1052.9	0.000
True vs Tchol	0.630	56.1	0.000

		b-value	standard error	p-value	r^2
True vs treat and LDL-chol	treat	0.0135	0.006	0.032	0.98
	LDL-chol	0.891	0.003	0.000	
True vs treat and Tchol	treat	0.375	0.005	0.000	0.75
	Tchol	0.685	0.018	0.001	

True = true endpoint; treat = treatment modality (0 or 1 for placebo and active treatment); LDL-chol = surrogate endpoint LDL-cholesterol level; Tchol = surrogate endpoint total cholesterol level; r^2 = Pearson's correlation coefficient squared; b = regression coefficient.

total cholesterol levels are significant predictors of the true endpoint in the multiple regression model with respectively b = 0.891, se = 0.003, p = <0.0001 and b = 0.685, se = 0.018, p < 0.0001. However, the subtraction sum of the r^2-values is 0.75 – 0.25 = 0.50 for total cholesterol and 0.98 - 0.25 = 0.73 for LDL-cholesterol. If the surrogate endpoint is made the dependent variable instead of the true endpoint, then LDL-cholesterol performs better than does total cholesterol. For LDL-cholesterol $r^2 = 0.20$, p-values < 0.01, power approximately 80%; for total cholesterol $r^2 = 0.04$, p = 0.226.
We can conclude that in order to establish a powerful correlation between treatment modality and a surrogate endpoint (p < 0.01, power > 80%), the proportion variability in y explained by the surrogate endpoint should be close to 70% or more for accurate predictions.
A wrong method is to accept as a valid result a surrogate endpoint that is a significant determinant of the true endpoint but not of the treatment modality.

We should add that different regression models are more convenient for different data like logistic regression models for odds ratios and Cox regression for survival data, but that the approach, otherwise, is similar.

6. DISCUSSION

In trials using surrogate endpoints the sample sizes have to be based not only on the expected treatment efficacy but also on the validity of the surrogate marker used. A method for calculating adjusted samples sizes is given.

Binary surrogate endpoints can be validated by calculating sensitivity and specificity to predict the true endpoint. However, overall validity is hard to quantify using this dual approach. Instead, an overall validity can be expressed by the proportion of patients that have a true surrogate test, either positive or negative, which we called the overall validity level. Still other approaches to the validity of surrogate tests are the so-called positive and negative predictive values and likelihood ratios. Just like the overall validity level, these estimators adjust for numbers of differences in patients with and without the true endpoint, but unlike the overall validity level they do not answer what proportion of patients has a correct test. Riegelman[9] , recently, proposed as method for assessing validity of diagnostic tests the discriminant ability, defined as (sensitivity + specificity) / 2. Although this method avoids the dual approach to validity, it wrongly assumes equal importance and equal prevalence of true positive and true negatives, and does neither answer what proportion of the patients has a correct test. We, therefore, decided to use an overall validity level, expressed as the percentage of patients with a true surrogate test, either positive or negative. We calculated confidence intervals of this estimate in order to quantify the level of uncertainty involved in the trial results. If the 95% confidence interval of the data is entirely within a previously set interval of validity, then the surrogate marker can be validated for use in subsequent trials.

In case of continuous surrogate tests regression models are adequate for testing validity. Not only the association between surrogate and true endpoint must be accounted, but also the associations between either of these variables and the treatment modality to be tested. Interaction assessments are not necessary, if there are no clinical arguments for the presence of interaction. A surrogate test can be validated only, if the proportion of variability in the surrogate endpoint explains the true endpoint by 70% or more, because the power of the surrogate endpoint to determine the treatment effect is then about 80%. A wrong conclusion would be to accept adequate validity if the surrogate test is an independent determinant of the true endpoint but not of the treatment modality.

Validating surrogate endpoints can only be done in a trial where a sufficient number of patients reaches both the surrogate and the true endpoint. With mortality or major cardiovascular events as true endpoint large randomized trials with long term follow-up are needed for that purpose. Chen et al[8] proposed as an alternative a semi-large study with a validation and non-validation set of patients, but this approach is not really different from two separate studies in a single framework.

Another interesting alternative was recently proposed by Kassaï et al. They meta-analyzed multiple small studies, but their effort was limited by its post-hoc nature and the heterogeneity of the studies included.[10]

If the required sample size or length of follow-up cannot be accomplished, then validity testing of surrogates for true endpoints will be impossible. We will have to look for alternative research methods like looking for intermediate endpoints such as morbidity instead of mortality. We should add that there are additional problems with a true endpoint like mortality: (1) for estimating the effects of preventive medicine that is begun when subjects are middle-aged this endpoint will be statistically weak, because at such ages the background noise of mortality due to other conditions associated with senescence is high, (2) to individual patients low morbidity and high quality of life, generally, means more than does a few additional years of survival. Fortunately, in other research the true endpoint is very well possible, and the surrogate endpoint is pursued because of practical and costs considerations. This applies, e.g., to the example described in the above section. This paper was, particularly, written for the latter purpose. It is to be hoped that the paper will affect the validity of future clinical trials constructed with surrogate endpoints.

7. CONCLUSIONS

The International Conference of Harmonisation (ICH) Guideline E9 Statistics Principles for Clinical Trials recommends that surrogate endpoints in clinical trials be validated using either (1) the sensitivity-specificity approach or (2) regression analysis. The problem with (1) is that an overall level of validity is hard to give, and with (2) that a significant correlation between the surrogate and true endpoint is not enough to indicate that the surrogate is a valid predictor. This chapter provides for a nonmathematical readership procedures that avoid the above two problems.

1. Instead of the sensitivity-specificity approach we used an overall validity level, expressed as the percentage of patients with a true surrogate test, either positive or negative. We calculated confidence intervals of this estimate, and assessed whether they were entirely within the prespecified interval of validity. If so, the surrogate marker was validated for use in subsequent trials. 2. For validating continuous surrogate variables, regression analysis was used, accounting both the correlation between the surrogate and true endpoints, *and* the associations between these two variables and the treatment modalities to be tested. If the proportion of variability in the surrogate endpoint explained the true endpoint by 70% or more, the surrogate test was validated. A wrong conclusion would be here to accept validity if the surrogate endpoint was an independent determinant of the true endpoint, but not of the treatment modality. It is to be hoped that this paper will affect the validity of future clinical trials constructed with surrogate endpoints.

8. REFERENCES

1. Pratt Cm, Moye LA. The Cardiac Arrhythmias Suppression Trial. Circulation 1995; 91: 245-7.
2. Canner PL, Berg KG, Wenger NK, Stamler J, Friedman L, Prineas RJ, Friedewald F. Fifteen year mortality of the Coronary Drug Project. J Am Coll Cardiol 1986; 8: 1245-55.
3. Riggs P et al. Osteoporosis in postmenopausal women. N Engl J Med 1990; 32: 802-9.
4. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996; 125: 605-13.
5. Boissel JP, Collet Hc. Surrogate endpoints: a basis for a rational approach. Eur J Clin Pharmacol 1992; 43: 235-44.
6. Philips A, Haudiquet V. The International Conference of Harmonisation (ICH) Guideline E9, Statistics Principles for Clinical Trials. Stat Med 2003; 22: 1-11.
7. Cleophas TJ. Clinical trials: a new method for assessing accuracy of diagnostic tests. Clin Res Reg Aff 2005; 22: 93-101.
8. Chen SX, Leung DH, Qin J. Information recovery in a study with surrogate endpoints. J Am Stat Assoc 2003; 10: 7-18.
9. Riegelman RK. Studying a study and testing a test. Philadelphia, PA: Lippincott Williams & Wilkins, 2005.
10. Kassaï B, Shah NR, Leizorovicz A, Cucherat M, Gueyffier F, Boissel JP. The true treatment benefit is unpredictable in clinical trials using surrogate outcome measures with diagnostic tests. J Clin Epidemiol 2005; 58: 1042-51.

CHAPTER 39

METHODS FOR REPEATED MEASURES ANALYSIS

1. INTRODUCTION

Repeated measures of the same kind, like, for example, blood pressures obtained from a single subject at subsequent times, are different from single measures of separate subjects, because repeated measures in a single subject are generally more similar to one another than those obtained from entirely different subjects, and statistical analyses have to take this difference into account. For that purpose there are paired and unpaired t-tests, repeated and non-repeated measures analysis of variance (ANOVA) (Chapters 1 and 2), paired and unpaired tests for binary data (Chapter 3), and paired and unpaired non-parametric tests (Chapter 2). Paired data or repeated measures means that multiple observations are performed in a single subject with the advantage that less subjects are required for answering a scientific question. A special type of repeated measures are longitudinal data including times series and survival data. They have been discussed in the chapters 16 and 43.

The most important reason for writing this chapter is the fact that repeated measures are frequently analyzed inappropriately. A linear regression or unpaired t-test or non-repeated measures ANOVA using repeated values does not take into account the repeated nature of the data, and, is therefore, likely to overestimate the magnitude of differences in the data. For example, drug-elimination curves[1] and R2-like models[2] for predicting probabilities of events are usually assessed with linear regression in spite of the repeated measures character of the data.

The current chapter reviews methods for repeated measures of continuous data. Particular attention will be given to (1) summary measures and (2) repeated measures ANOVA with and (3) without between-subjects covariates.

2. SUMMARY MEASURES

It is appropriate when possible to use a summary estimate of repeated data. For example, the area under the curve of drug concentration-time curves is used in clinical pharmacology as an estimate of bioavailability of a drug. Also, maximal values, mean values, changes from baseline are applied. The disadvantage of these measures is, that they do not use the data fully, because they use summary measures instead, and, therefore, may lose precision, but, otherwise, they are unbiased, and can be used perfectly well.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 461–466.

3. REPEATED MEASURES ANOVA WITHOUT BETWEEN-SUBJECTS COVARIATES

Summary measures are impossible if we want to assess the differences between the separate measures.
The study in Table 1 shows that a repeated measure ANOVA can be performed in a sample size as small as 4 subjects. The study compares the effects of three different treatments to reduce vascular resistance and contains only 12 data. A condensed version of this example was already presented in

Table 1. Repeated measures ANOVA, effects of three treatments on vascular resistance (blood pressure / cardiac output)

Subject	treatment 1	treatment 2	treatment 3	SD^2
1	22.2	5.4	10.6	147.95
2	17.0	6.3	6.2	77.05
3	14.1	8.5	9.3	18.35
4	17.0	10.7	12.3	21.4
Treatment mean	17.58	7.73	9.60	

Grand mean = 11.63

$SS_{\text{within subj}} = 147.95 + 77.05 + \ldots$.
$SS_{\text{treatment}} = (17.58 - 11.63)^2 + (7.73 - 11.63)^2 + \ldots$.
$SS_{\text{residual}} = SS_{\text{within subject}} - SS_{\text{treatment}}$

SPSS[1] statistical software: command: analyze; general linear model; repeated measurements.
Mauchly's test of sphericity chi-square = 2.07, 2 dfs, p = 0.355. No inequality of variance is in the data.

	SS	dfs	mean square	F	p-value
Within subject	127.2	1	127.2	127.2 / 7.0 = 18.2	0.024
Residual	20.964	3	7.0		

SS = sum of squares; dfs= degrees of freedom equals

Chapter 2. According the data-analysis there is a significant difference between the three treatments with F = 18.2 and p = 0.024. In order to assess the appropriateness of the linear assumption of the model a quadratic relationship between indicator and outcome variables, but the F-value and thus p-value were smaller (11.01 and p = 0.045), and, so the linear assumption seems appropriate. Between subjects differences are assumed not to influence the treatment-comparisons and are therefore not taken into account in the data analysis.

4. REPEATED MEASURES ANOVA WITH BETWEEN-SUBJECTS COVARIATES

In the study of Table 2 an example is given of a study where both repeated and non-repeated factors are combined. Three treatment modalities for the treatment of exercise tachycardias are assessed in both male and female subjects of different age classes, 20-30, 30-40, and 40-60 years of age. The variable 1 gives the age class (respectively 1, 2, and 3). The variable 5 gives the genders (respectively 1 and 2).

Table 2. Repeated measures ANOVA with between-subjects covariates, data-file of 54 subjects, the variables are explained in the text (the SPSS file uses commas instead of dots, VAR = variable).

Subject	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5
1.	1,00	112,00	166,00	215,00	1,00
2.	1,00	111,00	166,00	225,00	1,00
3.	1,00	89,00	132,00	189,00	1,00
4.	1,00	95,00	134,00	186,00	2,00
5.	1,00	66,00	109,00	150,00	2,00
6.	1,00	69,00	119,00	177,00	2,00
7.	2,00	125,00	177,00	241,00	1,00
8.	2,00	85,00	117,00	186,00	1,00
9.	2,00	97,00	137,00	185,00	1,00
10.	2,00	93,00	151,00	217,00	2,00
11.	2,00	77,00	122,00	178,00	2,00
12.	2,00	78,00	119,00	173,00	2,00
13.	3,00	81,00	134,00	205,00	1,00
14.	3,00	88,00	133,00	180,00	1,00
15.	3,00	88,00	157,00	224,00	1,00
16.	3,00	58,00	99,00	131,00	2,00
17.	3,00	85,00	132,00	186,00	2,00
18.	3,00	78,00	110,00	164,00	2,00
19.	1,00	112,00	166,00	215,00	1,00
20.	1,00	111,00	166,00	225,00	1,00
21.	1,00	89,00	132,00	189,00	1,00
22.	1,00	95,00	134,00	186,00	2,00
23.	1,00	66,00	109,00	150,00	2,00
24.	1,00	69,00	119,00	177,00	2,00
25.	2,00	125,00	177,00	241,00	1,00
26.	2,00	85,00	117,00	186,00	1,00
27.	2,00	97,00	137,00	185,00	1,00
28.	2,00	93,00	151,00	217,00	2,00
29.	2,00	77,00	122,00	178,00	2,00
30.	2,00	78,00	119,00	173,00	2,00
31.	3,00	81,00	134,00	205,00	1,00
32.	3,00	88,00	133,00	180,00	1,00
33.	3,00	88,00	157,00	224,00	1,00
34.	3,00	58,00	99,00	131,00	2,00
35.	3,00	85,00	132,00	186,00	2,00
36.	3,00	78,00	110,00	164,00	2,00
37.	1,00	112,00	166,00	215,00	1,00
38.	1,00	111,00	166,00	225,00	1,00
39.	1,00	89,00	132,00	189,00	1,00
40.	1,00	95,00	134,00	186,00	2,00
41.	1,00	66,00	109,00	150,00	2,00
42.	1,00	69,00	119,00	177,00	2,00

43.	2,00	125,00	177,00	241,00	1,00
44.	2,00	85,00	117,00	186,00	1,00
45.	2,00	97,00	137,00	185,00	1,00
46.	2,00	93,00	151,00	217,00	2,00
47.	2,00	77,00	122,00	178,00	2,00
48.	2,00	78,00	119,00	173,00	2,00
49.	3,00	81,00	134,00	205,00	1,00
50.	3,00	88,00	133,00	180,00	1,00
51.	3,00	88,00	157,00	224,00	1,00
52.	3,00	58,00	99,00	131,00	2,00
53.	3,00	85,00	132,00	186,00	2,00
54.	3,00	78,00	110,00	164,00	2,00

SPSS statistical software[3]: command: analyze; general linear model; repeated measurements.
Mauchly's test of sphericity chi-square = 30.7, 2 dfs, $p < 0.0001$. Inequality of variance in the data cannot be excluded.
Therefore, the Greenhouse-Geisser adjustment is used.

	SS	dfs	mean square	F	p-value
1.Within-subjects	281916.3	1.35	208537.9	2814.9	0.000
2. VAR 1 (age class)	4681.0	2	2340.5	3.045	0.057
3. VAR 5 (gender)	26373.0	1	26373.0	34.3	0.000
4. VAR 1 x VAR 5	2155.4	2	1077.7	1.402	0.256
5. Within-subjects x VAR 1	241.7	2.70	89.4	1.206	0.313
6. Within-subjects x VAR 5	1034.8	1.35	765.4	10.332	0.001
7. Within-subjects x VAR 1 x VAR 5	1481.9	2.70	548.1	7.398	0.000

SS = sum of squares; dfs = degrees of freedom

The variables 2, 3, and 4 give the exercise heart rate during treatment with respectively high dose, low dose and very low dose beta-blocker in beats / min. The study tries to answer 7 research questions:

1. Does treatment modality influence exercise heart rate?
2. Do subjects from different age classes have different heart rates?
3. Do males have different heart rates from females?
4. Does the pattern of differences between pulse rates for the age class groups change between the genders?
5. …….. for the age class groups change between the treatment modalities?
6. …… ……..for genders change between the treatment modalities?
7. …….. ……for treatment modalities change in a subgroup with a particular gender and age class?

The above research questions given in research terms are the following:

1. Is there a within-subjects main effect?

2. …….a between-subjects main gender effect?
3. …….a between-subjects main age class effect?
4. …….a between-subjects interaction
5-7……..a within-subjects by between-subjects interaction effect?

Because the test for non-equality of the variances can not be rejected, the usual ANOVA-model is inappropriate. Either an adjusted ANOVA, e.g., Greenhouse-Geisser adjusted univariate repeated measures ANOVA, or Multivariate Analysis of Variance (MANOVA) has to be applied. MANOVA is, conceptually, different, because it assumes multiple outcome variables instead of a single-one-with-multiple-levels. SPSS produces an analysis of both approaches. In our example the results of the two were virtually the same. According the data-analysis there was a significant difference between the three treatments with $F = 2814.9$ and $p = 0.000$. In order to assess the appropriateness of the linear assumption of the model a quadratic relationship between indicator and outcome variables, but the F-value and thus p-value were smaller, and, so the linear assumption seems appropriate.
It can be hard to interpret the results of interactions. For example, the above 7th question is confirmed: there is a significant interaction between treatment modality and gender and age class. Averages of the subgroups can be examined in order to understand what is going on (Figure 2). The Figure 2 shows that gender differences

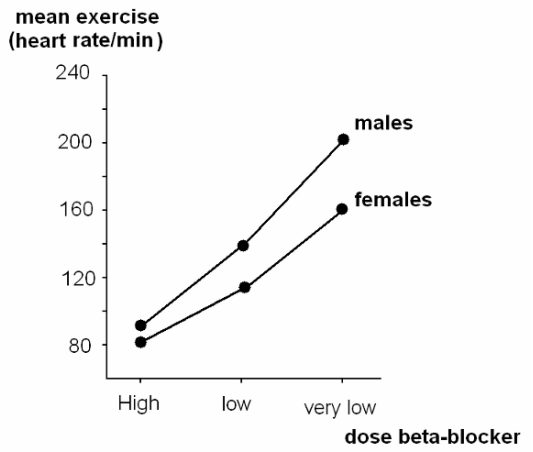


Figure 2. Gender differences do not remain the same but seem to increase with subsequent treatment modalities. There is interaction between treatment modality and gender. However, this is only true in the younger, but not in the older age classes.

do not remain the same but seem to increase with subsequent treatment modalities. There is, obviously, interaction between treatment modality and gender. However, this is only true in the younger, but not in the older age classes.

Generally, a significant interaction is a disaster for a comparative study of different treatment modalities, because an overall comparison of the treatment modalities becomes meaningless. In the given example the magnitude of the interaction is limited, and an overall treatment differences can still be observed (Figure 2). The overall result can be reported, at least, in a qualitative manner.

5. CONCLUSIONS

1. Repeated measures in a single subject are generally more similar to one another than data obtained from entirely different subjects, and statistical analyses have to take this into account.
2. It is appropriate when possible to use a summary estimate if the repeated data. For example, the area under the curve of drug time-concentration and time-efficacy curves, maximal values, mean values, change from baseline.
3. In parallel-group studies the level of statistical significance of between-subjects differences is usually assessed. In repeated measures within-subjects differences are usually assessed for that purpose. The general linear model available in SAS, SPSS and other statistical software programs provides repeated measures ANOVA appropriate for that purpose.
4. Repeated measures ANOVA can also adequately include subgroup factors like gender differences and age class differences into the analysis.
5. Like any type of ANOVA equality of variances and linearity in the data have to be checked. Most software programs routinely do so, and present alternative approaches in case these requirements can not be satisfied..

6. REFERENCES

1. Benet LZ, Kroetz DL, Sheiner LB. Pharmacokinetics. In: Goodman and Gilman's Pharmacologic Basis Therapeutics, 9th Edition, McGraw-Hill, New York, 1997 pp 5-27.
2. Hoetker G. The use of logit and probit models in strategic management research: critical issues. Strat Mgmt J 2007; 28: 331-43.
3. SPSS Statistical Software 13.0, Chicago , IL, USA, 2000.

CHAPTER 40

ADVANCED ANALYSIS OF VARIANCE, RANDOM EFFECTS AND MIXED EFFECTS MODELS

1. INTRODUCTION

In clinical trials it is common to assume a fixed effects research model. This means that the patients selected for a specific treatment are assumed to be homogeneous and have the same true quantitative effect and that the differences observed are residual, meaning that they are caused by inherent variability in biological processes, rather than some hidden subgroup property. If, however, we have reasons to believe that certain patients due to co-morbidity, co-medication, age or other factors will respond differently from others, then the spread in the data is caused not only by the residual effect but also by between patient differences due to some subgroup property. It may even be safe to routinely treat any patient effect as a random effect, unless there are good arguments no to do so. Random effects research models require a statistical approach different from that of fixed effects models.[1-3]

With the fixed effects model the treatment differences are tested against the residual error, otherwise called the standard error. With the random effects models the treatment effects may be influenced not only by the residual effect but also by some unexpected, otherwise called random, factor, and so the treatment should no longer be tested against the residual effect. Because both residual and random effect constitute a much larger amount of uncertainty in the data, the treatment effect has to be tested against both of them.[4,5]

Random effects models have been used in several studies recently published.[6-12] They are a very interesting class of models, but even a partial understanding is fairly difficult to achieve. This chapter was written to explain random effects models in analysis of variance and to give examples of studies qualifying for them.

2. EXAMPLE 1, A SIMPLE EXAMPLE OF A RANDOM EFFECTS MODEL

In a particular study the data may be different from one assessing doctor to the other due to differences in personality, education, or other doctor-factors. The example in

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 467–477.

Table 1. Example 1, a simple example of a random effects model

assessing doctor no	1	2	3	4	5	
patient	5.8	6.0	6.3	6.4	5.7	
	5.1	6.1	5.5	6.4	5.9	
	5.7	6.6	5.7	6.5	6.5	
	5.9	6.5	6.0	6.1	6.3	
	5.6	5.9	6.1	6.6	6.2	
	5.4	5.9	6.2	5.9	6.4	
	5.3	6.4	5.8	6.7	6.0	
	5.2	6.3	5.6	6.0	6.3	+
	44.0	49.7	47.2	50.6	49.3	Add-up sum = 240.8

The computations are:

$$\text{SS total} = \sum (y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 1{,}455.94 - \frac{(240.8)^2}{40} = 6.32$$

$$\text{SS between} = \frac{(44.0)^2 + \ldots (49.3)^2}{8} - \frac{(240.8)^2}{40} = 3.48$$

SS within = SS total –SS between = 6.32 – 3.48 = 2.84

Source	SS	df	MS	F	p-value
Between Assessors	3.48	5-1=4	0.87	0.87/0.08 = 10.72	<0.01
Within Assessors	2.84	40-5=35	0.08		

SS = sum of squares; df = degree of freedom; MS = mean square; F = test statistic for F-test

Table 1 gives the data of a study where a random sample of 5 doctors assess 8 different patients each. The data consist of individual health scores per patient. This example was modified from an example used by Hays.[13]

The scientific question is: “are the differences between the doctors larger than could happen by chance”. We have no prior theory that one or two particular assessing doctors will produce higher health scores than the rest, but rather expect that in the population of assessing doctors at large there may be heterogeneity for whatever reason. This means that a random effects model applies to this situation. We test whether between doctor variability compared to within doctor variability is large. Table 1 also gives the results of the analysis. For 4 and 35 degrees of freedom the F-test exceeds the F of 3.25, and the hypothesis of no difference between the

doctors is rejected. Indeed, there is a significant difference between the doctors. More in general the conclusion of this result should be that we can expect differences within any random sample of assessing doctors.
The calculations for a fixed effects model analysis of these data would produce the same result. The inference from it would, however, be entirely different: we would conclude that while we do not know anything about the population of assessing doctors at large, we definitely found a significant difference within this particular set of assessing doctors. It should be emphasized that the calculations for a fixed effect and a random effect are similar, but the interpretation is different. The choice between the fixed and random effect model is dependent on the statistical question. Both situations may be encountered in real life.

3. EXAMPLE 2, A RANDOM INTERACTION EFFECT BETWEEN STUDY AND TREATMENT EFFICACY

In clinical trials the observed differences between treatment modalities are compared to the differences caused by residual effects otherwise called noise. In studies with unexpected subgroup effects, this method is not appropriate and the increased variability in the data due to the subgroup effect has to be accounted. Random effects models are adequate for that purpose. An example is given underneath.

Table 2. Example 2, a random interaction effect between the study number and treatment efficacy

	Verapamil	Metoprolol	
Study 1	52	28	
	48	35	
	43	34	
	50	32	
	43	34	
	44	27	
	46	31	
	46	27	
	43	29	
	49	25	
	464	302	766
Study 2	38	43	
	42	34	
	42	33	
	35	42	
	33	41	
	38	37	
	39	37	
	34	40	
	33	36	

34	35	
368	378	746
832	680	

The computations are:

$$\text{SS total} = 52^2 + 48^2 + \ldots..35^2 - \frac{(52+ 48+ +\ldots35)^2}{40} = 1750.4$$

$$\text{SS treat by gender} = \frac{464^2 + \ldots 378^2}{10} - \frac{(52+ 48+ +\ldots35)^2}{40} = 1327.2$$

SS residual = SS total – SS treat by gender = 423.2

$$\text{SS rows} = \frac{766^2 + 746^2}{20} - \frac{(52+ 48+ +\ldots35)^2}{40} = 10.0 \text{ (= SS gender)}$$

$$\text{SS columns} = \frac{832^2 + 680^2}{20} - \frac{(52+ 48+ +\ldots35)^2}{40} = 577.6 \text{ (= SS treatment)}$$

SS interaction = SS treat by gender – SS rows – SS columns = 1327.2 – 10.0 – 577.6 = 739.6

Fixed effects analysis of variance

Source	SS	df	MS	F	p-value
Rows (study effect)	10.0	1			
Columns (treatment effect)	577.6	1	577.6	577.6/11.76 = 49.1	<0.0001
Interaction	739.6	1	739.6	739.6/11.76 = 62.9	<0.0001
Residual	423.2	36	11.76		
Total					

Random effects analysis of variance

Source	SS	df	MS	F	p-value
Rows (study effect)	10.0	1			
Columns (treatment effect)	577.6	1	577.6	577.6/739.6 = 0.781	ns
Interaction	739.6	1	739.6	739.6/11.76 = 62.9	<0.0001
Residual	423.2	36	11.76		
Total					

SS = sum of squares; df = degree of freedom; MS = mean square; F = test statistic for F-test; ns = not significant

The effects of two compounds on the numbers of episodes of paroxysmal atrial fibrillation is assessed in two rather similar parallel-group trials of 20 patients each. For the purpose of power the two studies are analyzed simultaneously. The data are given in Table 2. Overall metoprolol scores better than verapamil, but this is only true for the patients in study-1. There seems to be interaction between the study number and the treatment efficacy. The data are entered in the SPSS Software

program[14] commanding: statistics, general linear model, univariate. Choose as dependent variable numbers of episodes of paroxysmal atrial fibrillation, and as independent variables (1) treatment modality and (2) study number. The software program enables to treat the independent variables either as fixed or random variable. In the Table 2 are the results of the two assessments. If study-number is treated as a fixed effects variable, both treatment effect and interaction effect are compared to the residual effect. With 1 and 36 degrees of freedom the F-tests exceed the F of 5.57. Both a significant treatment effect and interaction effect is in the data. If treatments have different efficacies across studies, then an overall effect is not relevant anymore since the treatment effects cannot be interpreted independently of the interaction effect. The treatment efficacy of the treatment modalities is determined not only by the treatment modality but also by the study number. The information given by the random effect model is more adequate. The interaction effect is compared to the residual effect. With 1 and 36 degrees of freedom the F-test exceeds the F of 5.57. Subsequently, the treatment effect is compared not to the residual effect but rather to the interaction effect. With 1 and 1 degrees of freedom an F-value of 648 is required. The hypothesis of no treatment effect cannot be rejected. thus, a significant interaction exists and the overall treatment efficacy is not significant anymore. This result is obtained because the difference in the data due to different treatments is not compared with the residual differences but rather with the differences due to the interaction (which in this model includes the residual differences).

4. EXAMPLE 3, A RANDOM INTERACTION EFFECT BETWEEN HEALTH CENTER AND TREATMENT EFFICACY

Table 3. Example 3, a random interaction effect between the health center and treatment efficacy

Treatment	Verapamil	Metoprolol	Isosorbide mononitrate	Total
	number of attacks per patient			
Health center				
1	4	10	10	
	6	9	10 +	
	10	19	20	49
2	5	9	11	
	7	11	10 +	
	12	20	21	53
3	4	11	10	
	7	12	13 +	
	11	23	23	57

4	9	6	11	
	10	8	11 +	
	19	14	22	55
5	12	7	12	
	12	7	13 +	
	24	14	25	63
6	11	7	14	
	12	8	13 +	
	23	15	27	65
total	99	105	138	342

The computations are:

$$\text{SS total} = 4^2 + \ldots + 13^2 - \frac{(342)^2}{36} = 245$$

$$\text{SS ab} = \frac{10^2 + \ldots + 27^2}{2} - \frac{(342)^2}{36} = 224$$

SS error = SS total – SS ab
= 245 – 224 = 21

$$\text{SS columns} = \frac{99^2 + 105^2 + 138^2}{12} - \frac{342^2}{36} = 73.5$$

$$\text{SS rows} = \frac{49 + \ldots + 65^2}{6} - \frac{342^2}{36} = 30.67$$

SS interaction = SS ab – SS rows – SS columns
= 224 – 30.67 – 73.5 = 119.8

Fixed effects analysis of variance

Source	SS	df	MS	F	p-value
Rows (center effect)	30.67	6-1=5			
Columns (treatment effect)	73.5	3-1=2	36.75	36.75 /1.17= 31.49	<0.0001
Interaction(treatment x center)	119.8	2x5=10	11.98	11.98 /1.17= 10.24	<0.0001
Residual	21	18x(2-1)=18	1.17		
Total	245	35			

Random effects analysis of variance

Source	SS	df	MS	F	p-value
Rows (center effect)	30.67	6-1=5			
Columns (treatment effect)	73.5	3-1=2	36.75	36.75 /11.98= 3.07	ns
Interaction(treatment x center)	119.8	2x5=10	11.98	11.98 /1.17= 10.24	<0.0001

Residual	21	18x(2-1)=18	1.17
Total	245	35	

SS = sum of squares; df = degrees of freedom; MS = mean square; F = test statistic for F-test; ns = not significant

The effect of three compounds on the frequency of anginal attacks in patients with stable angina pectoris is assessed in a three group parallel-group study (Table 3). Current cardiovascular trials of new treatments often include patients from multiple health centers, national or international. Differences between centers may affect local results. We might say these data are at risk of interaction between centers and treatment efficacy. Patients were randomly selected in 6 health centers, 6 patients per center, and every patient was given one treatment at random, and so in each center two patients were given one of the three treatments.

When looking into the data we observe something special and unexpected. Metoprolol performs well in groups 4-6, i.e., better than in groups 1-3, and better than verapamil. This is unexpected, and may be due to interaction between the efficacy of treatment and the presence of particular health centers. There may be something about the combination of a particular health center with a particular treatment that accounts for differences in the data. For the analysis, as given in Table 3, SPSS statistical software[14] is used again using the commands: statistics, general linear model, univariate. The numbers of anginal attacks are the dependent variable, dependent variables are (1) treatment modalities and (2) health center. If health center is treated as a fixed effect variable, again both treatment effect and interaction effect are compared to the residual effect. With respectively 2 vs 18 and 10 vs 18 degrees of freedom the F-values of 4.46 and 2.77 are exceeded. Both a significant treatment effect and interaction effect is in the data. In multiple health centers we may have multiple treatment effects. The random effects method is more appropriate. With health center as random independent variable the analysis shows that with 10 vs 18 degrees of freedom the F-value of 2.77 is exceeded. A significant interaction exists. Subsequently, the treatment effect is tested against the interaction effect. With 2 and 10 degrees of freedom an F of 5.46 is required for significance, so that the hypothesis of no treatment effect cannot be rejected. The overall treatment efficacy is not significant anymore. This result is, like in the above example, obtained, because the difference in the data due to different treatments is not compared with the residual differences but rather with the differences due to the interaction. The following inference is adequate. Within the health centers, treatment differences apparently exist. Perhaps the capacity of a treatment to produce a certain result in a given patient depends on his/her health center background. Explanations include environmental factors like social and ethnic factors, investigator factors.

5. EXAMPLE 4, A RANDOM EFFECTS MODEL FOR POST-HOC ANALYSIS OF NEGATIVE CROSSOVER TRIALS

Table 4. Example 4, a random effects model for crossover trials

	Treatment 1	treatment 2	treatment 3	treatment 4	sd^2
Patient no.					
1	49	48	49	45	..
2	47	46	46	43	..
3	46	47	47	44	
4	47	45	45	45	
5	48	49	49	48	
6	47	44	45	46	
7	41	44	41	40	
8	46	45	43	45	
9	43	41	44	40	
10	47	45	46	45	
11	46	45	45	42	
12	45	40	40	40	
	552	540	540	528	Add-up sum = 2160

Fixed effects analysis of variance

Source	SS	df	MS	F	p-value
Between-patients	81.4	12-1=11			
Within-patients	226.6	3x12=36	6.29		
treatment	24.0	4-1=3	8	8/6.14 = 1.30	ns
residual	202.6	3x11=33	6.14		
Total	308.0	47			

Random effects analysis of variance

Source	SS	df	MS	F	p-value
Within-patients	226.6	12-1=11			
Treatments	24.0	4-1=3	8	8/1.74 = 4.60	<0.05
Subjects x treatments	57.5	3x11=33	1.74		
Total	308.0	47			

SS = sum of squares; df = degrees of freedom; MS = mean square; F = test statistic for F-test; sd = standard deviation; ns = not significant

In a crossover study different treatments are assessed in one and the same subject. An example of four treatments is given in Table 4. This example was also modified from an example used by Hays [1988]. A real difference between the treatments is expected and this is tested by comparing the observed differences between the treatment with the residual error, estimated from the subtraction of the sum of

squares (SS) within-patients minus the SS treatment. Obviously, none of the treatments produced an effect significantly different from that of another treatment. If a difference is not established like in this example, this may be due to random subgroup effects. In some patients one or more treatments may outperform the others, while in other patients other treatments may do so.
Because the study result was, thus, negative, we perform a post-hoc random effects analysis, testing treatment effect against treatments x patients interaction (Table 4, SPSS statistical software[14]; command: mixed model, linear). This assessment shows that the four treatments appear to be having different effects to different subsets. Some patients seem to respond better than the others to one or more treatments. This may due to personal factors like a genetic characteristic, a societal and/or developmental factor etc. Note that, although in the first analysis the SS within-patients has 36 degrees of freedom (number of patients x (number of treatment modalities -1)), in the second analysis it only has 11 degrees of freedom (number of patients -1). This is, because in the latter analysis the factors defining the treatment effects are considered to be fixed, while the subjects are viewed as randomly sampled.

6. DISCUSSION

In this chapter research models are discussed that account for variables with random rather than fixed effects. These models are often called type 2 models if they include random exposure variables and type 3 models if they include both fixed and random exposure variables. E.g., the example 1 in this paper gives a type 2 model while the last three examples are type 3 models, otherwise called mixed effects models. Another example of mixed effects models is the non-linear mixed effects (non-mem) modeling, increasingly for the development of pharmacokinetic parameters.[7,9,11,12.15] It is a program for nonlinear regression modeling that makes use of analysis of variance methods similar to those described in this paper.
The work-up of the advanced research models is sometimes largely the same as that of simple research models. But, inferences made are quite different. All inferences made under the simple model mostly concern means and differences between means. In contrast, the inferences made using advanced models deal with variances, and involve small differences between subsets of patients or between individual patients. This type of analysis of variance answers questions like: do differences between assessors, between classrooms, between institutions, or between subjects contribute to the overall variability in the data?
We should consider some limitations of the methods. If the experimenter chooses the wrong model, he/she may suffer from a loss of power. Also the standards of homogeneity / heterogeneity in the data must be taken seriously. The patients in the subsets should not be sort of alike, rather they should be exactly alike on the variable to be assessed. Often this assumption can not be adequately met, raising the risk of a biased interpretation of the data.
The random effects research models enable to assess the entire sample for the presence of possible differences between subgroups without need to, actually, split

the data into subgroups. This very point is a major argument in their favor. Also they are, of course, more appropriate if variables can be assumed to be random rather than fixed. A potential disadvantage is that the sensitivity to detect a significant difference in the data is generally somewhat reduced as explained in this paper. However, the reduction of sensitivity should not be regarded as a disadvantage, but rather an advantage, since the chance to make a correct conclusion is increased. Data should be analyzed according to the correct procedure, not according to the procedure that gives the largest chance to demonstrate a significant difference.

Only the simplest examples have been given in the present paper. The Internet provides an overwhelming body of information on the advanced research models including the type 2 and 3 research models as discussed here. E.g., the Google data system provides 495,000 references for explanatory texts on this subject. This illustrates the enormous attention currently given to these upcoming techniques. Yet in clinical research these models are little known. We hope that this paper will stimulate clinical investigators to more often apply them.

7. CONCLUSIONS

In clinical trials a fixed effects research model assumes that the patients selected for a specific treatment have the same true quantitative effect and that the differences observed are residual error. If, however, we have reasons to believe that certain patients respond differently from others, then the spread in the data is caused not only by the residual error but also by between patient differences. The latter situation requires a random effects model. This chapter explains random effects models in analysis of variance and to give examples of studies qualifying for them.

1. If in a particular study the data are believed to be different from one assessing doctor to the other, and if we have no prior theory that one or two assessing doctors produced the highest scores, but rather expect there may be heterogeneity in the population of doctors at large, then a random effect model will be appropriate. For that purpose between doctor variability is compared to within doctor variability.
2. If the data of two separate studies of the same new treatment are analyzed simultaneously, it will be safe to consider an interaction effect between the study number and treatment efficacy. If the interaction is significant, a random effects model with the study number as random variable, will be adequate. For that purpose the treatment effect is tested against the interaction effect.
3. In a multi-center study the data are at risk of interaction between centers and treatment efficacy. If this interaction is significant, a random effects model with the health center as random variable, will be adequate. The treatment effect is tested not against residual but against the interaction.
4. If in a crossover study a treatment difference is not observed, this may be due to random subgroup effects. A post-hoc random effects model, with patients effect as random variable, testing the treatment effect against treatments x patients interaction, will be appropriate.

Random effects research models enable the assessment of an entire sample of data for subgroup differences without need to split the data into subgroups. Clinical investigators are generally hardly aware of this possibility and, therefore, wrongly assess random effects as fixed effects leading to a biased interpretation of the data.

8. REFERENCES

1. Anonymous. Distinguishing between random and fixed variables, effects and coefficients. Newson, USP 656 Winter 2006, p1-3.
2. Campbell MJ. Random effects models. In: Statistics ar square two, second edition. Editor Campbell MJ. Blckwell Publishing, BMJ Books, Oxford UK, 2006, pp 67-83.
3. Gao S. Special models for sampling survey. In: Advanced Medical Statistics, first edition. Editors Lu Y, Fang J. World Scientific, New Jersey, 2003, pp 685 -709.
4. Anonymous Variance components and mixed models. http://www.statsoft.com/textbook/stvarcom.html
5. Anonymous. Random effects models.Wikipedia, the free encyclopedia. en.wikipedia.org/wiki/random-effects_models.html
6. Brier ME, Aronoff GR. Application of artificial neural networks to clinical pharmacology. Int J Clin Pharmacol Ther. 1996 Nov; 34: 510-4.
7. Dalla Costa T, Nolting A, Rand K, Derendorf H. Pharmacokinetic -pharmacodynamic modelling of the in vitro antiinfective effect of piperacillin -tazobactam combinations. Int J Clin Pharmacol Ther. 1997 Oct; 35: 426-33.
8. Mahmood I. Center specificity in the limited sampling model (LSM): can the LSM developed from healthy subjects be extended to disease states? Int J Clin Pharmacol Ther. 2003 Nov; 41: 517-23.
9. Meibohm B, Derendorf H. Basic concepts of pharmacokinetic / pharmacodynamic (PK/PD) modelling. Int J Clin Pharmacol Ther. 1997 Oct; 35: 401-13. Review.
10. Lima JJ, Beasley BN, Parker RB, Johnson JA. A pharmacodynamic model of the effects of controlled-onset extended-release verapamil on 24-hour ambulatory blood pressure. Int J Clin Pharmacol Ther. 2005 Apr;43(4): 187-94.
11. Lotsch J, Kobal G, Geisslinger G. Programming of a flexible computer simulation to visualize pharmacokinetic-pharmacodynamic models. Int J Clin Pharmacol Ther. 2004 Jan; 42: 15-22.
12. Mueck W, Becka M, Kubitza D, Voith B, Zuehlsdorf M. Population model of the pharmacokinetics and pharmacodynamics of rivaroxaban--an oral, direct factor xa inhibitor--in healthy subjects. Int J Clin Pharmacol Ther. 2007 Jun; 45: 335-44.
13. Hays WL. Random effects and mixed models, chapter 13. In: Statistics, 4th edition, Holt, Rhinehart and Winnston Inc, Chicago, 1988, pp 479-543.
14. SPSS Statistical Software. http://www.spss.com
15. Boeckman AJ, Sheiner LB, Beal SL. NONMEM User's Guide. San Francisco: NONMEM Project Group, University of California, 1992, Book.

CHAPTER 41

MONTE CARLO METHODS

1. INTRODUCTION

For more than a century statistical tests based on Gaussian curves have been applied in clinical research, like t-tests, chi-square tests, analysis of variance and most regression analyses methods. Current cardiovascular trials often make use of convenience samples and small samples that do not follow Gaussian curves. This raises the risk of false negative results.[1] Alternatively, samples can be analyzed using the Monte Carlo method. Basically, the Monte Carlo method uses random numbers from your own study rather than assumed Gaussian curves to assess the data. It was invented by the physicist Stanislaw Ulam[2] in the post-world-war-II era, and was called by him after the city of the roulette, because roulette is a simple generator of random numbers. The Monte Carlo method is, actually, very general: all it requires is the use of random numbers. It allows you to examine complex issues more easily than advanced mathematical methods, including integrals and matrix algebra. It is currently found in everything from economics to regulating flow of traffic to quantum mechanics. In cardiovascular research the Monte Carlo method has been recently applied, for example, for the analysis of brachytherapy data[3], computer tomographic images[4], pharmacological data[5], and observational data.[6] Overall, however, the Monte Carlo method is little used in cardiovascular research. This is a pity given the great potential of this relatively new method. This paper was written to elucidate its principle and gives some real data-examples for a non-mathematical readership. The body of ongoing cardiovascular research is huge, and cardiovascular investigators tend to perform basic statistics without the help from a statistician. This paper was written for their benefit. It is to be hoped that the paper will stimulate them to use the Monte Carlo method more often, particularly in case of convenience samples and small numbers.

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 479–486.

2. PRINCIPLES OF THE MONTE CARLO METHOD EXPLAINED FROM A DARTBOARD TO ASSESS THE SIZE OF Π

The basics of the Monte Carlo method was explained by Woller[7] using a dartboard for the purpose of assessing the size of π. Figure 1 simply pictures one quadrant of

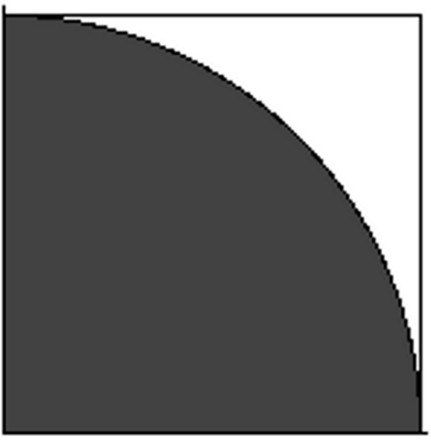


Figure 1 A very poor dart player is assumed to have equal chances of throwing darts inside and outside the circle area, a situation which simulates throwing darts randomly (r = radius of the dartboard), (with permission from the author).[11]

a circle. We assume that a very poor dart player throwing darts at it produces the same result as that obtained by throwing darts randomly at the figure. In this case the number of darts in the circle quadrant is proportional to the area of that part of the Figure. This would mean:
number darts in circle quadrant / total number darts = area circle quadrant / total area of graph.

High school geometry told us (r = radius of circle):

$$\text{area circle quadrant / total area of graph} = \tfrac{1}{4}\,\pi\,r^2 / r^2 = \tfrac{1}{4}\,\pi.$$

If at random a dart lands somewhere inside the graph, the ratio of hits in the circle quadrant will be one-fourth the value of π. Throwing many darts can thus be used as a method to assess the size of π.

π = 4 x (area circle quadrant / total area of graph)
= 4 x (number darts in circle quadrant / total number darts)

However, if you actually use this type of experiment for the assessment of π, you will observe that it will take a large number of throws to obtain a reliable value of π… well over 1,000. Yet, it is a straightforward alternative to the advanced mathematical methods commonly used to solve the problem. In clinical data analysis the Monte Carlo method can serve a similar purpose.

3. THE MONTE CARLO METHOD FOR ANALYZING CONTINUOUS DATA

Table 1. The bootstrap method is a data based simulation process for statistical inference. The basic idea is randomly picking up a patient from a given sample while replacing the picked-up patient so that the sample from which to choose remains unchanged.

Original data		bootstrap 1		bootstrap 2	
LDL-cholesterol (mmol/l)					
group1	group 2	group1	group 2	group 1	group 2
1. 3.99	10. 3.18	1. 3.99	10. 3.18	1. 3.99	10. 3.18
2. 3.79	11. 2.84	1. 3.99	10. 3.18	2. 3.79	11. 2.84
3. 3.60	12. 2.90	3. 3.60	12. 2.90	2. 3.79	12. 2.90
4. 3.73	13. 3.27	5. 3.21	14. 3.85	2. 3.79	12. 2.90
5. 3.21	14. 3.85	6. 3.60	15. 3.52	4. 3.73	14. 3.85
6. 3.60	15. 3.52	8. 3.61	15. 3.52	5. 3.21	15. 3.52
7. 4.08	16. 3.23	8. 3.61	15. 3.52	7. 4.08	16. 3.23
8. 3.61	17. 2.76	9. 3.81	16. 3.23	7. 4.08	18. 3.60
9. 3.81	18. 3.60	9. 3.81	17. 2.76	8. 3.61	18. 3.60
median group 1 = 3.73		median group 1 = 3.61		median group 1 = 3.79	
median group 2 = 3.23		median group 2 = 3.23		median group 2 = 3.23	
difference medians 0.50		difference medians = 0.38		difference medians = 0.55	

Table 1 gives an example of a parallel-group study assessing the effect of two cholesterol-reducing treatments on low density lipoprotein (LDL)-cholesterol. Small samples like those given here often do not follow a Gaussian curve. Non-parametric testing can restore a Gaussian curve. However, sometimes it is not good enough. SPSS Software[8] is helpful. In the main box for two-samples-non-parametric testing the possibility of a test for the adequacy of non-parametric testing is given: the Kolmogorov-Smirnov (KS) test. If you click the KS test and then ok, a p-value of 0.037 is given, indicating that the KS test is positive and that

non-parametric testing is, indeed, not good enough. If we, subsequently, click "exact" in the main dialog box, another dialog box will occur. It gives you the possibility to use either an exact test or the Monte Carlo method. Exact tests make use of rank numbers, i.e., all individual results are given a rank-number in ascending order, and these ranks are added up to determine the exact chance of finding an overall result in your data. The problem with rank testing is that it rapidly runs into numerical problems that even modern computers have difficulty to solve. SPSS statistical software is helpful regarding this problem. When clicking "exact" for the second time, the program will highlight the text "exact method will be used instead of Monte Carlo when computational limits allow". You should set your computational time limits, e.g. 5 or 10 minutes, and the program will automatically use the Monte Carlo method if your requested time limits can not be met. In our particular example, the exact test required only 2 minutes and produced a p-value of 0.010 while the Monte Carlo method took less than a few seconds and produced a p-value of 0.011, virtually the same.

For the continuous data as given in this example a special type of Monte Carlo method is used, called the bootstrap method.[9] The name bootstrap derives from the saying "pull yourself up by your bootstraps" meaning that you can continue what you are doing but in a much faster way. It works essentially as follows. A patient is randomly picked up from the given samples while replacing the picked-up patient so that the sample from which to choose remains unchanged. In mathematical terms this method of sampling is called "sampling with replacement". By doing so, we can produce random samples from the original data.

The procedure is illustrated in Table 1. In the first random sample observation-1 was picked up twice, while observations-2 and -4 were not. We repeat the procedure a large number of times, and record the difference between medians every time. To derive reliable confidence intervals, at least 1000 repetitions are required. Medians rather than means are used, because with extreme high (or low) values the mean value is less representative of the data average than the median. The null hypothesis is, that no real difference exists between the data from group 1 and 2. This null hypothesis is rejected if the median of group 1 is larger than that of group 2 at least 95% of the times. All of the 1000 differences between the medians as calculated from the bootstraps lay between –0.12 and 1.09, while 98.9% of the differences lay between 0.00 and 0.844. In this example the bootstrap medians from group1 are indeed larger than those from group 2 over 95% of the times. We, therefore, reject the null hypothesis at $p < (1\text{-}0.95)$ or $p < 0.05$. There is a significant difference between the groups at $p < 0.05$ ($p = 0.011$ to be precise).

4. THE MONTE CARLO METHOD FOR ANALYZING PROPORTIONAL DATA

Table 2. 2 x 2 contingency table of a population-based cohort study assessing the effect of a prophylactive treatment on the numbers of cardiac events.

	Patients with an event		Patients without an event		proportion
Observed	cell 1	5	cell 2	995	5/1000
Expected from target population	cell 3	10	cell 4	990	10/1000

Table 2 gives an example of a population-based cohort study assessing the effect of a prophylactic treatment on the number of cardiac events. For proportional data, including fractions and percentages, the chi-square test is a standard method of analysis. The data are usually displayed in 2 x 2 contingency tables (Table 2). However, the cells of a 2 x 2 contingency table must not be too small, 5-10 patients are required in each of its four cells. If smaller, the Fisher's exact test is an alternative, but with proportions from large groups computational problems will rapidly arise, because it uses faculties: "995 faculty" = 995 ! = 995 x 994 x 993 x 992 x 991 ...etc. These kinds of calculations are time-consuming even for modern computers. The SPSS program is helpful again. If you click "exact" in the main dialog box, then another dialog box will occur. You should set your computational time limits, e.g. 5 minutes, and the program will automatically use the Monte Carlo method if your requested time limits will exceed.

The Monte Carlo method to calculate is, then, less labour-intensive, and works essentially as follows. The question is answered: are the observed cells significantly different from the cells expected from the population data base. If the proportion of patients-with-event in the target population can be expected to be 10 out of 1000, then the ratio 10:1000 will be observed most frequently when randomly sampling from such a target population. How great is the chance of finding a ratio 5:1000 if 10:1000 is to be expected. This chance will be small. We can randomly pick up a patient 1000 times from a sample of 1000, 10 of which have the code "event", and 990 of which have the code "no-event", while replacing the picked-up patient, so that the sample from which to choose remains unchanged (sampling with replacement).

	Patients with an event
The first 1000 patients chosen produced the following result	9
Second 1000 patients	10
Third 1000 patients	8
…..	4
….	...
.....	..
Thousandth 1000 patients	12

The null hypothesis is, that the difference between observed and expected is due to chance rather than a statistically significant effect. This null hypothesis is rejected at $p<0.05$ if in 95% of the times the expected number of patients with an event is larger than the observed number 5. In this example all of the 1000 pick-up procedures produced results between 4 and 16, while 95% of them were between 6 and 14, which is consistently larger than the observed number 5. We can, therefore, reject the null hypothesis at $p<0.05$.

5. DISCUSSION

The Monte Carlo method is a scientifically safe alternative approach to data analysis. It, essentially, derives confidence intervals from the data without prior assessment of the type of frequency distribution. Other advantages of this method include:

- It does not require equal standard deviations of groups in paired or parallel-group treatment comparisons.
- It is often less time-consuming than exact methods.

Disadvantages include:

- Although less laborious than many standard methods, it is still rather laborious without a computer.
-Samples must not be very small. With a sample of say four, there are only 16 distinct re-samplings equally likely. The median of such a sample will take one of the four sample values. This is not a very strong basis for constructing a 95% confidence interval. The above examples show that rather small samples are generally no problem. The p-value produced by the Monte Carlo method in the first study comparing 9 versus 9 patients was only slightly larger than the p-value produced by the exact test, with p-values of 0.011 and 0.010 respectively.

Nowadays Monte Carlo methods can be carried out with computer programs for statistical analyses, like SPSS, S-plus, StatsDirect, StatXact, SAS etc.[8,10-16] The current paper gives only the simplest examples of the Monte Carlo method for the analysis of clinical data. Several books have been written providing more complex models.[17-19] However, all of the models are based on the same simple principle. We do hope that this paper will strengthen the awareness the great potential of the Monte Carlo method for the analysis of research data.

6. CONCLUSIONS

For more than a century statistical tests based on Gaussian curves have been applied in clinical research. Current cardiovascular trials often make use of convenience samples and small samples that do not follow Gaussian curves. This raises the risk of false negative results. This chapter elucidates the Monte Carlo method as an alternative method for the assessment of such data.

The Monte Carlo method derives confidence intervals from the data without prior assumption about the presence of Gaussian curves in the data. For 2-parallel-groups studies with continuous data the basic idea is to produce multiple random samples from your own 2 parallel groups. If in at least 95 % of these random samples the first group scores better than the second, then a statistically significant difference between the two groups will be accepted at $p < 0.05$.

Also for population-based cohort studies with proportional data multiple random samples can be produced from your own observed data. If in at least 95% of these random samples the expected proportion exceeds the observed proportion, then a statistically significant difference between the observed and expected data will be accepted at $p < 0.05$.

Advantages of the Monte Carlo method include:

- It does not depend upon Gaussian curves.
- It is less time-consuming than many standard methods.

A disadvantages is that, although less time-consuming than many standard methods, it is still rather laborious without a computer. We do hope that this paper will strengthen the awareness of the Monte Carlo method as an often more reliable alternative for analysis of cardiovascular research.

7. REFERENCES

1. Zwinderman AH, Cleophas TJ, Van Ouwerkerk B. Clinical trials do not use random samples anymore. Clin Res Reg Affairs 2006; 23: 85-95.
2. Ulam N, Ulam S. The Monte Carlo method. J Am Stat Assoc 1949; 44: 335-41.
3. Vieira JW, Lima FR, Kramer R. A Monte Carlo approach to calculate dose distribution around the lineal brachytherapy sources. Cell Mol Biol 2002; 48: 445-50.
4. Haidekker YG. Trans-illumination optical tomography of tissue-engineered blood vessels: a Monte Carlo simulation. Appl Opt 2005; 44: 4265-71.
5. Upton RN, Ludbrook GL. Pharmacokinetic-pharmacodynamic modelling of the cardiovascular effects of drugs-method development and application in sheep. BMC Pharmacol 2005; 5: 1471-81.
6. Nijhuis RL, Stijnen T, Peeters A, Witteman JC, Hofman A, Hunink MG. Apparent and internal validity of a Monte Carlo-Markow model for cardiovascular disease in a cohort follow-up study. Med Dec Making 2006; 26: 134-44.

7. Woller J. The basics of Monte Carlo Simulation (1996). http://www.chem.unl.edu/zeng/joy/mclab/mcintro.html
8. SPSS Statistical Software. http://www.spss.com
9. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall, 1993, New York, USA.
10. S-plus.http://www.mathsoft.com/splus
11. StatsDirect. http://www.camcode.com
12. StatXact. http://www.cytel.com/products/statxact/statact1.html
13. True Epistat. http://ic.net/~biomware/biohp2te.htm
14. BUGS y WinBUGS. http://www.mrc-bsu.cam.ac.uk/bugs
15. R. http://cran.r-project.org
16. SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas
17. Gardner MJ, Altman DG. Confidence in analysis. Statistical software edited by BMJ, 1989, ISBN 07279 0222-9.
18. Shao J, Tu D. The jackknife and bootstrap. Springer-Verlag, New York, 1995.
19. Fishman GS. Monte Carlo, concepts, algorithms and applications. Springer -Verlag, New York, 1996.

CHAPTER 42

PHYSICIANS' DAILY LIFE AND THE SCIENTIFIC METHOD

1. INTRODUCTION

Physicians' daily life largely consists of routine, with little need for discussion. However, there are questions physicians simply do not know the answer of. Some will look for the opinions of their colleagues or the experts in the field. Others will try and find a way out by guessing what might be the best solution. The benefit of the doubt doctrine[1] is often used as a justification for unproven treatment decisions, and, if things went wrong, another justification is the expression: clinical medicine is an error-ridden activity.[2] So far, few physicians have followed a different approach, the scientific method. The scientific method is, in a nutshell: reformulate your question into a hypothesis and try to test this hypothesis against control observations. In clinical settings this approach is not impossible, but rarely applied by physicians, despite their lengthy education in evidence based medicine, which is almost entirely based on the scientific method. This paper was written to give simple examples of how the scientific method can be implied in a physician's daily life, and to explain its advantages and limitations. We do hope that this paper will stimulate physicians to more often apply the scientific method for a better outline of their patients' best possible treatment options.

2. EXAMPLE OF UNANSWERED QUESTIONS OF A PHYSICIAN DURING A SINGLE BUSY DAY

We assumed the numbers of unanswered questions in the physicians' daily life would be large. But just to get of impression, one of the authors of this paper (TC) recorded all of the unanswered answers he asked himself during a single busy day. Excluding the questions with uncertain but generally accepted answers, he included 9 questions.

During the hospital rounds 8.00-12.00 hours.

1. Do I continue, stop or change antibiotics with fever relapse after 7 days treatment?

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 487–493.

2. Do I prescribe a secondary prevention of a venous thrombosis for 3, 6 months or permanently?
3. Should I stop anticoagulant treatment or continue with a hemorrhagic complication in a patient with an acute lung embolia?
4. Is the rise in falling out of bed lately real or due to chance?
5. Do I perform a liver biopsy or wait and see with liver function disturbance without obvious cause?

During the outpatient clinic 13.00-17.00 hours.

6. Do I prescribe aspirin, hydroxy-carbamide or wait and see in a patient with a thrombocytosis of $800x10^{12}$ /l over 6 months?
7. Are fundic gland polyps much more common in females than in males?

During the staff meeting 17.00-18.00 hours

8. Is the large number of physicians with burn out due to chance or the result of a local problem?
9. Is the rise in patients' letters of complaints a chance effect or a real effect to worry about?

Many of the above questions did not qualify for a simple statistical assessment, but others did. The actual assessments, that were very clarifying for our purposes, are given underneath.

3. HOW THE SCIENTIFIC METHOD CAN BE IMPLIED IN A PHYSICIAN'S DAILY LIFE

1.Falling out of bed

Falling out of bed is the prime cause of injury in hospitalized patients, and the prevention of it is a high priority and criterion for quality care.[3,4] If more patients fall out of bed than expected, a hospital department will put much energy in finding the cause and providing better prevention. If, however, the scores tend to rise, another approach is to first assess whether or not the rise is due to chance, because daily life is full of variations. To do so the numbers of events observed is compared the numbers of event in a sister department. The pocket calculator method is a straightforward method for that purpose.

	Patients with fall out of bed	patients without	
department 1	16 (a)	26 (b)	42 (a+b)
department 2	5 (c)	30 (d)	35 (c+d)
	21 (a+c)	56 (b+d)	77 (a+b+c+d)

Pocket calculator method:

$$\text{chi-square} = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = 5.456.$$

If the chi-square value is larger than 3.841, then a statistically significant difference between the two departments will be accepted at $p<0.05$. This would mean that in this example, indeed, the difference is larger than could be expected by chance and that a further examination of the measures to prevent fall out of bed is warranted.

2.Evaluation of fundic gland polyps

A physician has the impression that fundic gland polyps, an otherwise rather benign condition, are more common in females than it is in males. Instead of reporting this subjective finding, he decides to follow the next two months every patient in his program.

	patients with fundic gland polyps	patients without	
females	15 (a)	20 (b)	35 (a+b)
males	15 (c)	5 (d)	20 (c+d)
	30 (a+c)	25 (b+d)	55 (a+b+c+d)

Pocket calculator method:

$$\text{chi-square} = \frac{(ad-bc)^2\,(a+b+c+d)}{(ab)(c+d)(b+d)(a+c)} = 5.304$$

The calculated chi-square value is again larger than 3.841. The difference between males and females is significant at $p<0.05$. We can be for about 95% sure that the difference between the genders is real and not due to chance. The physician can report to his colleagues that the difference in genders is to be taken into account in future work-ups.

3.Physicians with a burn-out

Two partnerships of specialists have the intention to associate. However, during meetings, it was communicated that in one of the two partnerships there were three specialists with burn-out. The meeting decided not to consider this as chance finding, but requested a statistical analysis of this finding under the assumption that unknown factors in partnership 1 may place these specialists at an increased risk of obtaining a burn-out.

	physicians with burn out	without burn out

partnership 1 (a+b)	3 (a)	7 (b)	10
partnership 2 (c+d)	0 (c)	10 (d)	10
(a+b+c+d)	3 (a+c)	17(b+d)	20

pocket calculator method

$$\text{chi-square} = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = \frac{(30-0)2\,(20)}{10x10x17x3} = \frac{900x20}{\ldots\ldots\ldots} = 3.6$$

The chi-square value was between 2.706 and 3.841. This means that no significant difference between the two partnerships exists, but there is a trend to a difference at $p<0.10$. This was communicated back to the meeting and it was decided to disregard the trend. Ten years later no further case of burn-out had been observed.

4.Patients' letters of complaints

In a hospital the number of patients' letters of complaints was twice the number in the period before. The management was deeply worried and issued an in-depth analysis of possible causes. One junior manager recommended that prior to this laborious exercise it might be wise to first test whether the increase might be due to chance rather than a real effect.

	patients with letter of complaints	patients without	
year 2006 (a+b)	10 (a)	1000 (b)	1010
year 2005 (c+d)	5 (c)	1000 (d)	1005
(a+b+c+d)	15 (a+c)	2000 (b+d)	2015

$$\text{chi-square} = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} = 1.64..$$

The chi-square was smaller than 2.706, and so the difference could not be ascribed to any effect to worry about but rather to chance. No further analysis of the differences between 2006 and 2005 were performed.

There are of course many questions in physicians' daily life that are less straightforward and cannot be readily answered with a pocket calculator. E.g., the effects of subgroups and other covariates in a patient group will require t-tests, analyses of variance, likelihood ratio tests, and regression models. Fortunately, in the past 15 years user-friendly statistical software[5-12] and self-assessment

programs[13] have been developed that can help answering complex questions. Nowadays, many clinical investigators already use them without the help of a statistician. However, we as authors of this paper are not aware of any physician who is not involved in a research program and still assesses his/her every-day questions in the way illustrated in the above examples.

4. DISCUSSION

Since the era of Hippocrates, 500 years BC, physicians have had an ethical obligation not only to provide the best possible care for their patients but also to enhance health to the entire population. Statistical tests have been recognized to produce the best evidence you can get from your data, and physicians applying them thus serve their population in the best possible way. A second point is that most practicing physicians are not avid readers of clinical trials. The information of new treatments is, instead, often brought to them by the media, the pharmaceutical industry and even the patients. This is not necessarily a criticism of well-trained and hard-working doctors. The problem is that the language of the published reports is such that physicians are almost as lost as a layperson, particularly, when it comes to the core sections of the report, the statistical analysis and result sections. It follows that either the results are accepted with too little of scepticism or rejected with too much of it. Particularly, the former may take place if a pharmaceutical representative communicates overstated the benefits and understated the risks. Being actively involved in the scientific method is a strong antidote against these hazards. Also, physicians start better reading the published clinical research and understanding its strengths and limitations, and, most important, its implications to health.

Do we have guarantees that the result is true if statistically significant. No, but it is the best evidence from your data you can get. There is, of course, the chance of type I errors of finding an effect which is a non-effect. This chance is particularly large with multiple testing. Then there is the chance of a type II error of finding no effect where there is one. This chance is particularly large with small samples. Third, you may be mistaking because you can not predict with full confidence if your target population is older, younger, from a different gender, or from any other sampling distribution than that of your test sample. But there are more limitations with the application of the scientific method. Hard-working doctors tend to have a full agenda, and, usually, do not have the leisure to write a study protocol, and rewrite it several times as required by their institutions' ethic committees, and find it hard to complete an entire informed consent procedure. Instead, many study protocols, particularly, those of observational studies, do not necessarily require ethic approval and written informed consent. None of the examples given needed the latter. A subsequent limitation is the limited validity of the chi-square test with samples smaller than 5. A final limitation is the possible damage in the patient-doctor relationship sometimes attributed to scientific activities in a daily practice. Indeed, many patients may expect from their doctor a more personal relationship based on thrust and sympathy, and in addition the best possible treatment. Telling a

patient of the risks of being in a placebo-control group and thus receiving nothing for his condition is not a typical basis for thrust. We should add that observational studies in this context are more patient-friendly than clinical trials. At least in observational studies patients are not recruited for a randomized treatment, but rather treated following their voluntary clinical visits.

5. CONCLUSIONS

So far, few physicians have followed the scientific method for answering practical questions they simply do not know the answer of. The scientific method is, in a nutshell: reformulate your question into a hypothesis and try to test this hypothesis against control observations. This chapter gives simple examples of how the scientific method can be implied in a physician's daily life.

Of 9 unanswered daily questions, 4 qualified for simple statistical assessments, which were very clarifying for the physicians involved. Additional advantages of the scientific method include:

(1) since the scientific method has been recognized to produce the best evidence you can get from your observations, physicians applying it serve their population in the best possible way; (2) being actively involved in the scientific method is a strong antidote against the hazards of accepting published studies from others with too little of scepticism or rejecting them with too much of it.

Limitations of the scientific method include: (1) type I and II errors; (2) misinterpretations due to different frequency distributions, (3) lack of leisure time on the part of the physicians to write a study protocol, (4) the risk of a damaged patient-doctor relationship.

6. REFERENCES

1. Ordronaux J. The jurisprudence of medicine in relation to the law of contracts, torts and evidence. The Lawbook Exchange, LTD, 1869.
2. Paget MA, The unity of mistakes, a phenomenological interpretation of medical work. Contemporary Sociology 1990; 19: 118-9.
3. Lambert V. Improving safety, reducing use. FDA Consumer 1992; October issue pp 1-5.
4. Anonymous. Beds in hospital, nursing homes and home health care. Drugs & Health products. 2001; May issue pp 2-6.
5. SPSS Statistical Software. http://www.spss.com
6. S-plus.http://www.mathsoft.com/splus
7. StatsDirect. http://www.camcode.com
8. StatXact. http://www.cytel.com/products/statxact/statact1.html
9. True Epistat. http://ic.net/~biomware/biohp2te.htm
10. BUGS y WinBUGS. http://www.mrc-bsu.cam.ac.uk/bugs
11. R http://cran.r-project.org SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas

12. Cleophas TJ, Zwinderman AH, Cleophas TF. Statistics applied to clinical trials: self-assessment book. Ed by Cleophas TJ, Kluwer Academic Publishers, Boston,MA, 2002.

CHAPTER 43

CLINICAL TRIALS: SUPERIORITY-TESTING

1. INTRODUCTION

One of the flaws of modern statistics is that it can produce statistically significant results even if treatment effects are very small. E.g., a sub-analysis of the SOLVD study[1] found symptoms of angina pectoris in 85.3% of the patients on enalapril and in 82.5% of the patients on placebo, difference statistically significant at $p<0.01$. In a situation like this one has to question about the clinical relevance of the small difference. Another problem of clinical trials is that the statistics is increasingly complex, and that clinicians are at a loss to understand it. This is not, necessarily, a criticism of well-trained and hard-working doctors, but it does have a very dark side. Studies are, generally, accepted if the magic p-values are <0.05, and the disappointment about the small benefit to individual patients comes later. The problem is that a p-value of 0.05 means that the power of finding a true positive effect is only 50%, and, more important, the chance of not finding it is equally 50%. Such a result is hardly acceptable for reliable testing.

The objectives of the current study were (1) to give some examples of studies that have been published as unequivocally positive studies, although the treatment effects were substantially smaller than they were expected to be, (2) to introduce superiority-testing as a novel statistical approach avoiding the risk of statistically significant but clinically irrelevant results. Superiority-testing defines a priori in the protocol clinically relevant boundaries of superiority of the new treatment. If the 95% confidence interval of the study result is entirely within the boundary, then superiority is accepted, and we do not have to worry about the p-values anymore.

2. EXAMPLES OF STUDIES NOT MEETING THEIR EXPECTED POWERS

The Lancet publishes benchmark research. We extracted from recent volumes of the Lancet six original articles of controlled clinical trials that were reported as being positive studies, although they did not meet their expected power. The studies produced only 53 to 83 % of the statistical power expected, while the new treatments produced only 46 to 86 % of the magnitude of response expected (Table 1). E.g., in the PROSPER study[4] the new treatment only produced half the benefit

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 495–503.

Table 1. Discrepancies between expected and observed statistical powers and treatment efficacies of six controlled clinical trials recently published in the Lancet

Study	sample size	comparison	expect / observ power (%)	expect / observ effect size	observ p-value
1. PPP study[2]	4495	aspirin vs placebo	90/48	from 5.4% to 2.9% / absolute risk reduction from 2.8% to 2.0%	0.055*
2. Staedke SG[3]	400	amiodiaquine vs sulfadioxine-pyrimethamine	80/66	15% / 7% absolute reduction treatment failures	0.023#
3. PROSPER[4]	5804	statin vs placebo	92 / 73	20% / 10% relative reduction events	0.015
4. ESTEEM[5]	1883	ximelagatran vs warfarin	80 / 56	27 % / 22 % relative reduction events	0.036
5. Jochan D[6]	379	adjuvant chemother vs no	90 / 63	2.10 / 1.58 hazard ratios	0.020
6. Andrews DW[7]	333	stereotactic radiosurgery vs no	80 / 53	3.5 / 1.6 months of survival	0.040

vs = versus; expect = expected; observ = observed; chemother = chemotherapy; *composite endpoint, this study was yet reported as a positive study, because separate endpoints were significant at 0.035-0.049); #the largest difference of the 3 main endpoints, the other two were not significant.

that was expected (10 % instead of 20 % relative reduction in events). In the Andrews study[7] the new treatment produced less than half the benefit expected (an average of 1.6 instead of 3.5 months of survival). These results, although statistically significant at the $p<0.05$ level, may not unequivocally demonstrate clinical superiority, and may not be good enough for accepting the new treatment for general use.

3. HOW TO ASSESS CLINICAL SUPERIORITY

The PROSPER study[4] included 5804 patients to test whether in elderly pravastatin performed better than placebo in preventing cardiovascular morbidity / mortality. The sample size in this study was based on an expected statistical power of 92% to observe a relative reduction of events of 20% (absolute reduction of 3.2%) with an absolute risk of events of 16% at baseline. Statistical power can be best described as the chance of finding a significant effect in your data, if there is a real effect. It

means that the expected chance that any real effect in the data is not detected (the type II error) is only 8 %. However, it turned out that the relative reduction in events was only 10% (absolute reduction 1.6%).

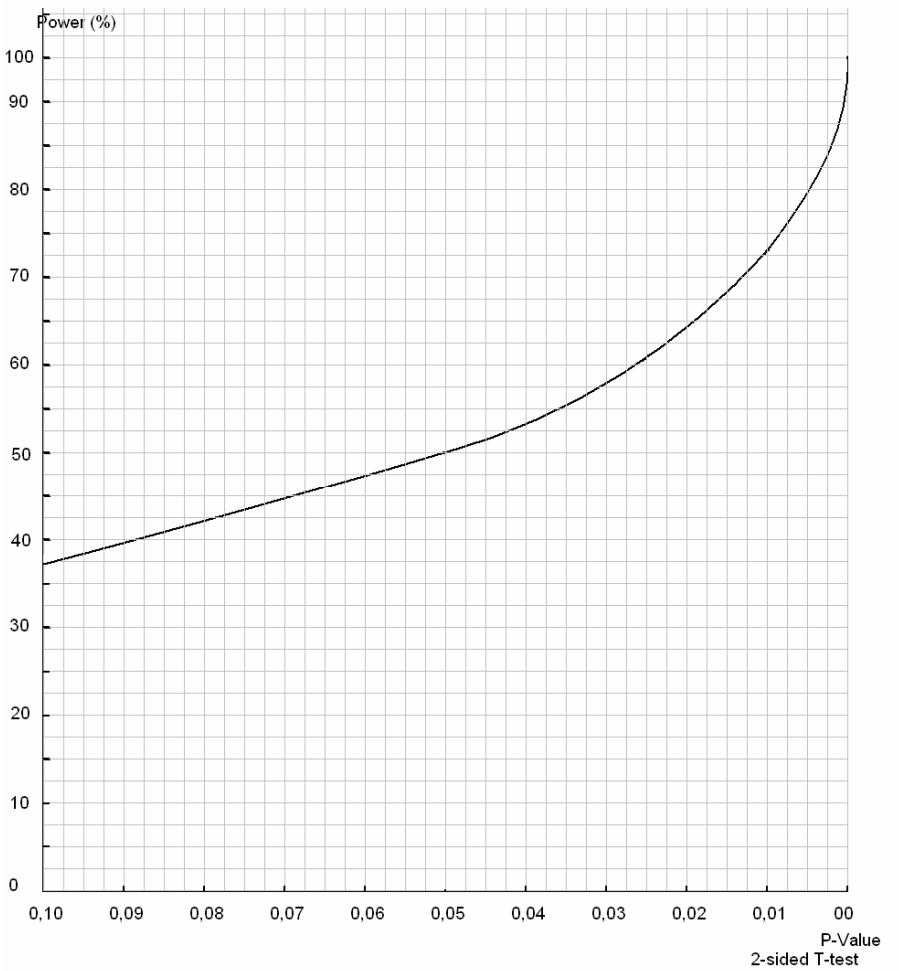


Figure 1: The relationship between power and p-values (two sided t-test with samples sizes > 200). The curve is approximately similar to the curve for the z-test (normal distribution).

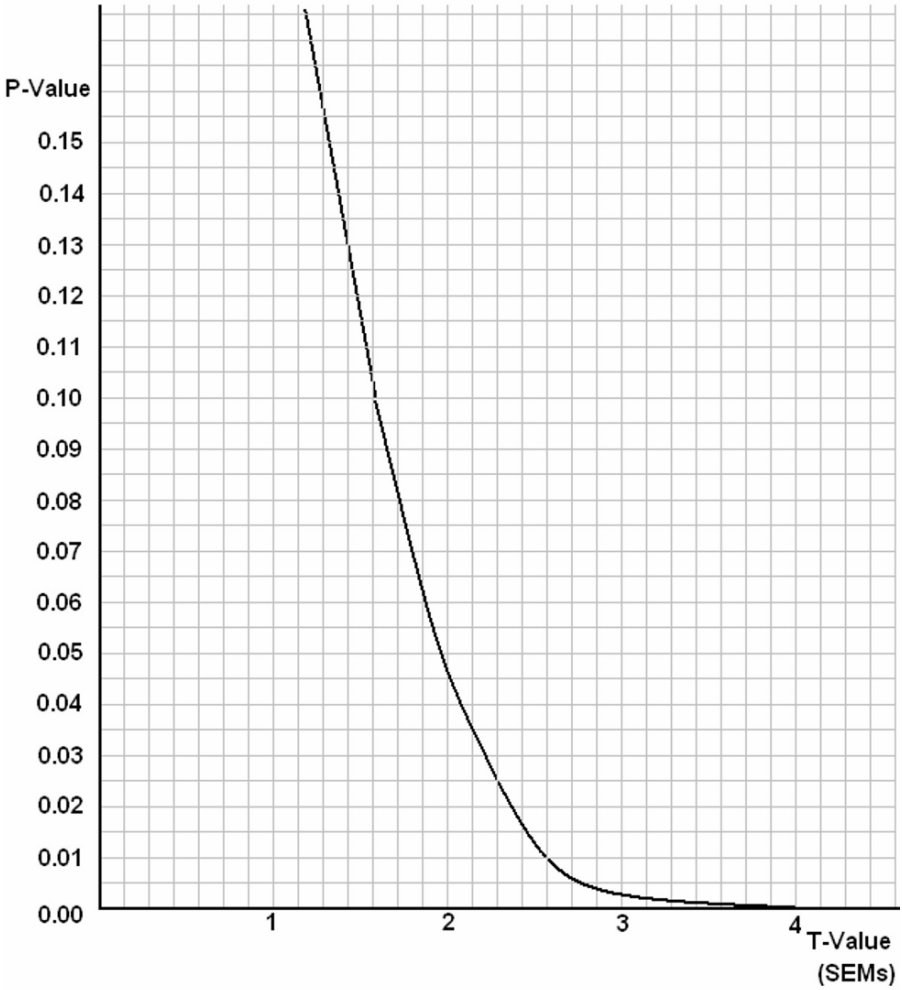


Figure 2: The relationship between p-values and t-values (two sided t-test with samples sizes > 200). The curve is approximately similar to the curve for the z-test (normal distribution).

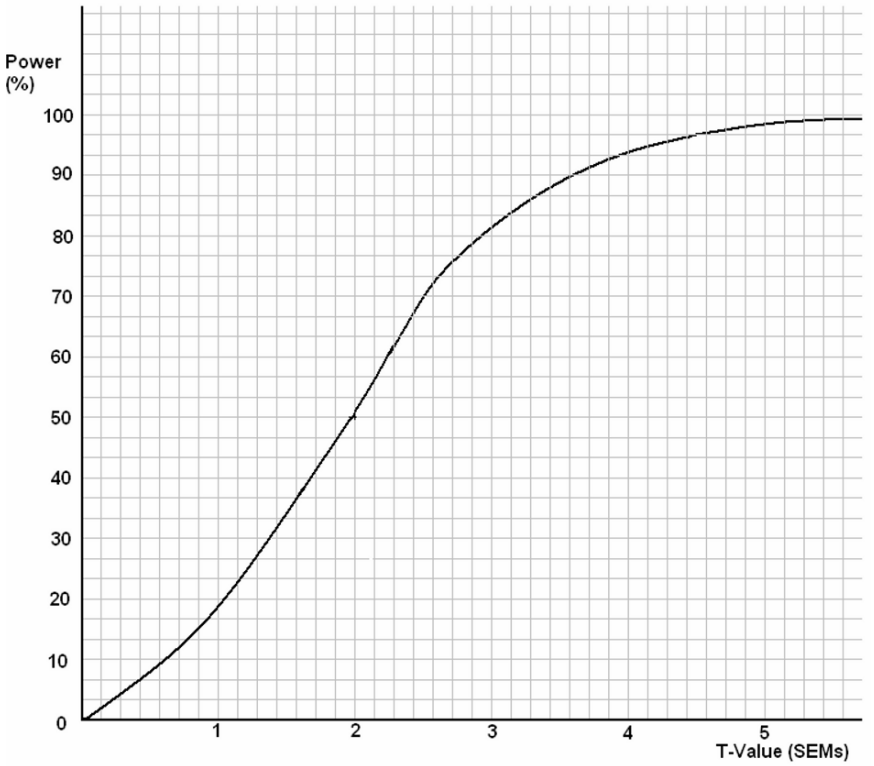


Figure 3: The relationship between power and t-values (two sided t-test with samples sizes > 200). The curve is approximately similar to the curve for the z-test (normal distribution).

Figures 1-3 give the relationships between statistical power, p-values, and t-values (two-sided t-tests with samples sizes > 200). The curves are approximately similar to the curves for the z-test (the test for normal distributions). T-values can be best interpreted as standardized measures of treatment efficacy; t-values larger than approximately 2 SEMs (standard errors of the mean) indicate that there is a significant effect at $p<0.05$ in the data. From the Figures 1-3 it can be extrapolated that the main endpoint result of the PROSPER study corresponded with a power of only 73%, instead of 92%, and, consequently, a type II error of 27% instead of the expected 8%. In spite of this disappointing result, the study reported that an unequivocal superiority of the new treatment had been demonstrated. However, the risk reduction observed was only half that expected, and the chance of a type II error of finding no difference next time, was 3.4 times that expected. This may not

be good enough a result for implementing the new treatment, particularly not, if potential adverse effects of the new treatment are taken into account.
Traditionally, in clinical trials a significant efficacy of a new treatment is accepted if the null-hypothesis of no treatment effect is rejected at p = 0.05, corresponding with a type II error of no less than 50%. This would mean for the PROSPER study a relative risk reduction of only approximately 5% (absolute risk reduction of 0.7%), which is not what one would call an impressive result. Instead of a p-value of 0.05 as cut-off criterion for demonstrating superiority a stricter criterion seems to be welcome. For that purpose an approach similar to that of equivalence-testing and non-inferiority-testing may be applied (Figure 4, upper two graphs). With

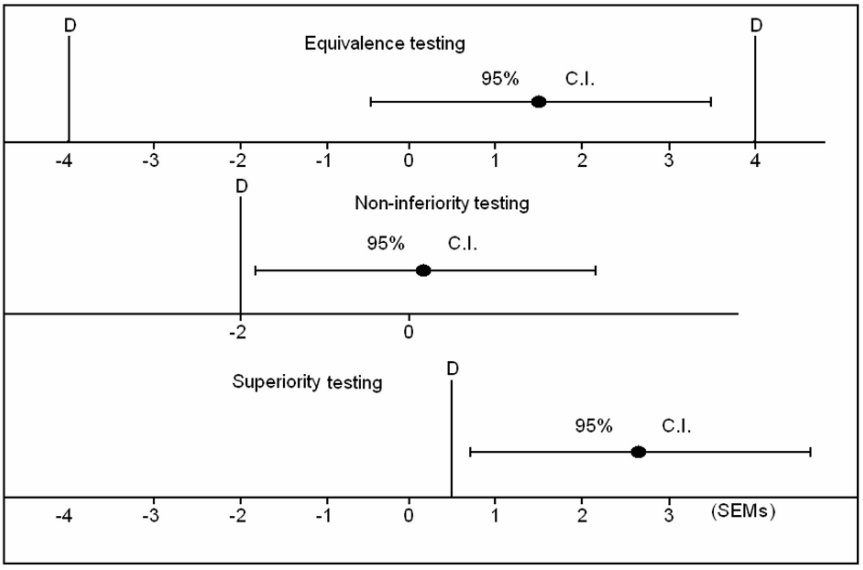


Figure 4. Examples of equivalence, non-inferiority- and superiority studies: any 95% confidence interval (C.I.) that does not cross the pre-specified range of equivalence, inferiority, or superiority as indicated by the D boundaries present the presence of equivalence, non-inferiority, or superiority respectively

equivalence / non-inferiority-testing we have prior arguments to assume little difference between the new treatment and control treatment, and we are more interested in similarity and non-inferiority of the new treatment versus control than in a statistically significant difference between the new treatment and control. A boundary of similarity or non-inferiority is a priori defined in the protocol. If the 95% confidence interval of the study turns out to be entirely within these boundaries, then similarity or non-inferiority is accepted.

Also for superiority-testing a prior boundary of superiority has to be defined in the study protocol. E.g., a boundary producing 10% less the power of the study's expected power could be chosen. Figure 5 shows what will happen if this boundary

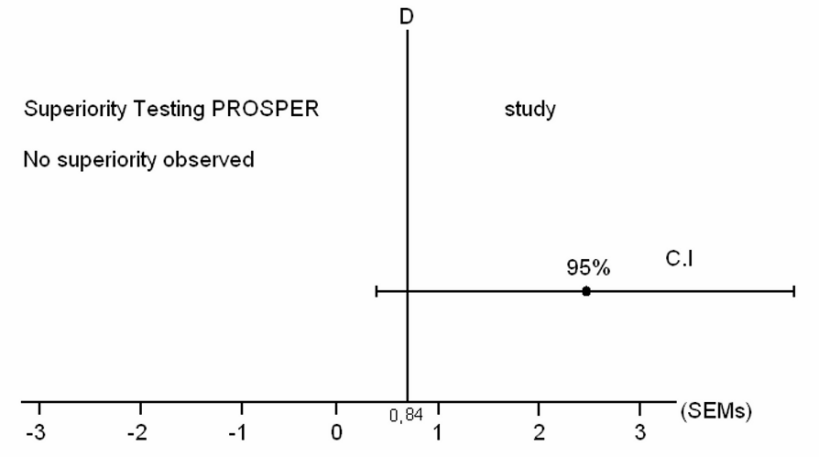


Figure 5. The 95% confidence interval (C.I.) of the PROSPER study crosses the D boundary. With the given D boundary this study is unable to demonstrate superiority.

is applied in the PROSPER study. The 95% confidence interval of the PROSPER study crosses this boundary and this means that the study is be unable to demonstrate superiority, and that the result is, therefore, negative.

4. DISCUSSION

Routinely replacing the assessment of p-values with superiority testing in comparative trials will have some advantages:

1. Studies with large type II errors will no longer be interpreted as positive studies, because small and irrelevant treatment efficacies, producing large type II errors, will no longer meet the criterion of superiority.
2. The general incentive to produce as small a p-value as possible, even if the study effect is very small will be gone. Specific methods for producing small p-values have been developed. They include the use of very large samples and composite endpoints. Very large samples will almost certainly show a statistically significant difference in the data, but this difference will be questionably clinically relevant. Composite endpoints produce small p-values, but are frequently complicated by large gradients in importance to patients result in misleading impressions of the

impact of treatment.[8] With superiority-testing as introduced in this paper, the p-values are no longer the criterion for a positive study.

3. P-values are, traditionally, applied for testing the null-hypotheses of no effect in the data. However, in current clinical trials the issue is not *any* effect in the data, but rather a *clinically relevant* effect or not. This latter question can never be answered by null-hypothesis testing, and requires a different approach. For that purpose clinical relevance has to be quantitatively defined, e.g. in the form boundaries of superiority, as introduced in the present paper.

We come to some important recommendations in this study. We recommend that investigators consider replacing testing null-hypotheses of comparative clinical trials with testing a priori defined boundaries of clinical superiority of new treatments. A similar approach is already common in equivalence-studies and non-inferiority-studies, but could very well be applied to the "normal" comparative studies usually performed for establishing the clinical superiority of one treatment over another. Nowadays, too many borderline significant studies are being reported as convincingly positive studies. This is a misleading practice as it produces overestimated expectations from new treatments.[9-12] Superiority-testing, as introduced in this paper, is a simple method to avoid this problem.

5. CONCLUSIONS

One of the flaws of modern statistics is that it can produce statistically significant results even if treatment effects are very small. The objective of the current chapter was (1) To give some examples of studies that have been published as unequivocally positive studies, although the treatment effects were substantially smaller than they were expected to be. (2) To introduce superiority-testing as a novel statistical approach avoiding the risk of statistically significant but clinically irrelevant results.

We extracted from recent volumes of the Lancet six original articles of controlled clinical trials that were reported as being positive studies, although they did not meet their expected power. The studies produced only 53 to 83 % of the statistical power expected, while the new treatments produced only 46 to 86 % of the magnitude of response expected. Instead of a p-value of 0.05 as cut-off criterion for demonstrating superiority a stricter criterion seems to be welcome. For that purpose, similar to equivalence-testing and non-inferiority-testing, prior boundaries of superiority have to be defined in the protocol. If the 95% interval of the study turns out to be entirely within these boundaries, then superiority is accepted.

Nowadays, too many borderline significant studies are being reported as convincingly positive studies. This is a misleading practice, as it produces overestimated expectations from new treatments. Superiority-testing, as introduced in this paper, is a simple method to avoid this problem.

6. REFERENCES

1. Yusuf S, Pepine CJ, Garces C, et al. Effect of enalapril on myocardial infarction and angina pectoris in patients with low ejection fraction. Lancet 1992; 340: 1173-8.
2. Collaborative Group of Primary Prevention Project. Low dose aspirin and vitamin E in people at cardiovascular risk: a randomised trial in general practice. Lancet 2001; 357: 89-75.
3. Staedke SG, Kanga MR, Dorsey G, et al. Amiodaquine, sulfadioxone/pyrimethamine and combination therapy for treatment of falciparum malaria in Kampala, Uganda: a randomised trial. Lancet 2001; 358: 368-74.
4. Shepherd J, Blauw GJ, Murphy MB, et al, on behalf of the PROSPER study group. Pravastatin in elderly individuals at risk of vascular disease: a randomised trial. Lancet 2002; 360: 1623-30.
5. Wallenstein L, Wilcox RG, Weaver WD, et al, for the ESTEEM investigators. Oral ximelagatran for secondary prophylaxis after myocardial infarction. Lancet 2003; 362; 789-97.
6. Jochan D, Richter A, Hofmann L, et al. Adjuvant autologous renal tumour cell vaccine and risk of tumor progression in patients with renal cell carcinoma. Lancet 2004; 362: 594-9.
7. Andrews DW, Scott CB, Speranto PW, et al. Whole brain irradiation therapy with or without stereotactic radiosurgery boost for patients with 1-3 brain metastases. Lancet 2004; 363: 1665-72.
8. Ferreira-Gonzalea I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007; 334: 786-8.
9. Horng S, Grudy C. Misunderstanding of clinical research. Ethics and Human Research 2003; 25: 11-6.
10. Fetting JH, Siminoff CA, Piantadosi S et al. effects of patients' expectations of benefit with standard breast cancer adjuvant therapy on participants in clinical trials. J Oncol 1990; 8: 1476-82.
12. Anonymous. Patients' demands for prescriptions in primary care. Br Med J 1995; 310: 1084-5.

CHAPTER 44

TREND-TESTING

1. INTRODUCTION

Some 15 years ago serious statistical analyses of cardiovascular trials were conducted by specialist statisticians using mainframe computers. Nowadays, there is ready access to statistical computing using personal computers, and this practice has changed boundaries between basic and more advanced statistical methods. Cardiovascular researchers, currently, perform basic statistics without professional help from a statistician, including t-tests and chi-square tests for two treatment comparisons. Current cardiovascular trials often involve more than two treatments or treatment modalities, e.g., dose-response and dose-finding trials, studies comparing multiple drugs from one class with different potencies, or different formulas from one drug with various bio-availabilities and other pharmacokinetic properties. In such situations small differences in efficacies are to be expected and we need, particularly, sensitive tests. A standard approach to the analysis of such data is multiple groups analysis of variance (ANOVA) and multiple groups chi-square tests, but a more sensitive, although so far little used, approach may be a trend-analysis. A trend means an association between the order of treatment and the magnitude of response. We should add that, within the context of a clinical trial, demonstrating trends, generally, provides more convincing evidence of causal treatment effects than do simple comparisons of treatment modalities.[1]
In the current paper we review methods for trend-analysis in cardiovascular trials that can be used by cardiologists without the support of a statistician. We also demonstrate that trend-tests may be more sensitive to demonstrate statistically significant treatment effects than do the standard methods for treatment comparisons.

2. BINARY DATA, THE CHI-SQUARE-TEST-FOR-TRENDS

For trend-analysis of binary data the chi-square-test–for-trends is adequate, although similar results can be obtained from logistic regression modeling. However, the former test is conceptually more straightforward and mathematically less complex. A real data example is given. In a hypertension trial responders were defined as patients with a blood pressure under 140/90 mm Hg. The data (Table 1) were first analyzed using multiple groups chi-square test, and this analysis produced a chi-square value of 3.872 with two degrees of freedom. According to the chi-square table this would mean, that this result is not statistically significant (p= 0.144). We have a negative study, that does not enable to conclude anything

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 505–510.

else than “no treatment differences in these data”. However, if we calculate the odds of responding (Table 1), we find incremental odds from treatment 0 to

Table 1. In a hypertension trial responders were defined as patients with a blood pressure under 140/90 mm Hg

	treatment 0	treatment 1	treatment 2	total
Number responders (d)	10	20	27	57 (O)
Number non-responders	15	19	15	49
Total number patients (n)	25	39	42	106 (T)
Odds of responding	0.67 (10/15)	1.11 (20/19)	1.80 (27/15)	

treatment 2, suggesting an association between the order of treatment and the magnitude of response, otherwise called a trend. The chi-square-test-for-trend can be used for assessment of this possible trend.

$\sum d = 10 \times 0 + 20 \times 1 + 27 \times 2 = 74$

$\sum n = 25 \times 0 + 39 \times 1 + 42 \times 2 = 123$

$\sum (n^2) = 25 \times 0 + 39 \times 1 + 42 \times 4 = 207$

$O = 57 \qquad T = 106 \qquad T\text{-}O = 49$

$U = \sum d - (O/T \times \sum n) = 74 - (57 / 106 \times 123) = 8.0$

$V = \frac{O\,(T\text{-}O)}{} \times (T \times \sum (n^2) - (\sum n)^2) = \frac{57 \times 49}{} \times (106 \times 207 - 123^2) = 16.079$

$T^2 \times (T\text{-}1) = 106^2 \times 105$

The chi-square-trend is calculated to be $U^2 / V = 3.980$ with one degree of freedom. According to the chi-square table this would mean, that we have a significant trend at $p < 0.05$. There is evidence that the higher the number of the treatment the more efficacious the treatment is. Particularly, if we have clinical arguments, like with increasing potencies of otherwise similar treatments, this result provides the evidence. Interestingly, the trend-test is significant in spite of a negative overall test for differences in the data. Obviously, a trend-test is sometimes more sensitive than a standard overall test to find differences in the data.

3. CONTINUOUS DATA, LINEAR-REGRESSION-TEST-FOR-TRENDS

For trend-analysis of continuous data linear-regression-modeling, is often used. As an example a hypertension trial with mean arterial blood pressures (MAPs) as efficacy variable is given (Table 2).

Table 2. In a hypertension trial mean arterial blood pressures (MAPs) after treatment were assessed as efficacy variable

	treatment 1	treatment 2	treatment 3
	number of patients		
	10	10	10
MAP (mm Hg)	122	118	115
	113	109	105
	131	127	125
	112	110	106
	132	126	124
	114	111	107
	130	125	123
	115	118	108
	129	124	115
	122	112	122
Mean	122	118	115
Standard deviation	8.08	7.15	8.08

First the data are assessed by an overall test. Multiple groups ANOVA is used for that purpose, and provides an F-value of 2.035 with 2 and 27 degrees of freedom. This result means that we have a p-value of 0.150, and, thus, no significant difference between the three groups of patients. Also, as expected, the largest difference between the mean MAP of treatments 1 and 3 are not significant in the unpaired t-test:

$$\text{t-value} = \frac{122-115}{\sqrt{(8.08^2 / 10 + 8.08^2 / 10)}} = 7 / 3.613 = 1.937$$

(with 20-2 = 18 degrees of freedom, $0.05 < p < 0.10$.

A linear-regression-test-for-trends using SPSS statistical software[2] produces the following results. We first enter the data or a data-file, e.g., from Excel.[3] Then we command: Statistics; Regression; Linear. Table 3 gives the results. The top-table

Table 3. Results of statistical analysis using SPSS software of the data from Table 2. The top-table gives the R-values, the middle-table tests with ANOVA whether R is significantly different from 0, the bottom-table provides the regression equation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,361[a]	,130	,099	7,64775

a. Predictors: (Constant), VAR00001

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	245,000	1	245,000	4,189	,050[a]
	Residual	1637,667	28	58,488		
	Total	1882,667	29			

a. Predictors: (Constant), VAR00001

b. Dependent Variable: VAR00002

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	125,333	3,694		33,927	,000
	VAR00001	-3,500	1,710	-,361	-2,047	,050

a. Dependent Variable: VAR00002

ANOVA = analysis of variance; df = degree of freedom; F= F-statistic; sig. = level of significance; R =correlation coefficient; B = regression coefficient; t= t-statistic.

calculates the correlation coefficient R and R^2 . The middle-table gives the result of testing with ANOVA whether R^2 is significantly different from 0. If $R^2 = 0$, the order of the treatment determines the MAPs no way, there is no trend. In our situation $R^2 = 0.13$, and, thus, 13% of the MAP results are determined by the differences in treatment modalities: there is a significant trend at $p = 0.05$. The bottom-table from Table 3 gives the regression equation that can be used to draw the best fit regression line for the data.

In spite of the negative ANOVA- and t-tests for treatment comparisons, there is a significant trend in the data. We are able to conclude that the order of the treatments is associated with the magnitude of efficacy. Like in the above example

of a binary variable, the trend-test was more sensitive than the standard tests to find differences in the data.

4. DISCUSSION

In this paper only trend-tests for parallel-group data are reviewed. Although not commonly used in cardiovascular trials and statistically somewhat more sophisticated, trend-tests are also available for repeated measurements in one subject or one group of patients. In the case of continuous data a linear mixed-model effect can be used, where the subjects are regarded as random variable, and the treatment as fixed effect variable. The presence of a significant treatment-by-subjects interaction is, then, considered as documented evidence of order-of-treatments-effect or trend. Analyses are available in SPSS and SAS.[2,4] Just like with parallel-group data a significant trend may be found in spite of a negative overall test for treatment differences. Repeated measurements with binary data are even less common in cardiovascular research, and statistical software for trends is also less generally available, but some methods are presented in SAS proc nlmixed.[4] To assess trends of odds ratios we can make use of the assumed normal distribution of the regression coefficients, the b-values. Log likelihood ratio tests with the b-value as variable can be used for the purpose.

Why is trend-analysis often more sensitive than standard testing? We should add that with two treatments a trend-test and a standard test provide identical results. This is, because we have equal degrees of freedom. However, with three or more treatment modalities the degrees of freedom with a standard analysis rapidly increase, while with trend-analysis they do not, giving rise to smaller p-values.

The limitations of trend-analysis have to be accounted. First, if there is no trend in the data, then the standard method of analysis may be more sensitive than the trend-test. So, standard tests should be performed in addition to the trend tests. Second, trend-testing assumes a linear trend in response in the data with subsequent treatments. This means that, with continuous data the means of the treatment groups increase linearly, and with binary data, the odds of responding increase exponentially (or the logarithms of the odds increase linearly). The linear effect is a simplifying assumption that should be checked. With only three categories as dependent variable, linearity is easy to check from a graph or even from the tables of the data. However, with multiple categories linearity checking is less straightforward, and special methods have to be used. Assuming a quadratic relationship between dependent and independent variable, and , then, performing a regression analysis is an adequate approach for that purpose, because the quadratic relationship is mathematically the simplest relationship that comes next to the linear relationship. If a better p-value is provided by the quadratic model, then this relationship should be pursued, and the linear relationship has to be abandoned.

Cardiovascular researchers, currently, perform basic statistics without professional help from a statistician, and current cardiovascular trials often involve more than two treatments or treatment modalities. Trend-tests may be more sensitive than standard methods for treatments comparisons. A trend means an association

between the order of treatment and the magnitude of response. The chi-square-test-for-trends and the linear-regression-test-for-trends are adequate for the analysis of parallel-group data. Although not commonly used in cardiovascular trials, trend-tests for repeated measurements in one subject or one group of patients are available in SPSS[2] , SAS[4], and other major statistical software programs.
Limitations of trend-testing include: (1) trend-testing may be less sensitive than standard tests if a trend in the data is lacking, (2) trends may not be linear. We recommend that trend-testing be included more routinely in cardiovascular trial-protocols in order to increase the sensitivity of data analysis of cardiovascular trials.

5. CONCLUSIONS

Cardiovascular researchers tend to perform basic statistics without professional help from a statistician, and current cardiovascular trials often involve more than two treatments or treatment modalities. Trend-tests may be more sensitive than standard methods for treatments comparisons. This chapter reviews methods for trend-analysis of parallel-group data from cardiovascular trials.

1. A trend means an association between the order of treatment and the magnitude of response.
2. The chi-square-test-for-trends and the linear-regression-test-for-trends are adequate for the analysis of parallel-group data.
3. Although not commonly used in cardiovascular trials, trend-tests for repeated measurements in one subject or one group of patients are available in SPSS , SAS, and other major statistical software programs.
4. Limitations of trend-testing include: (1) trend-tests may be less sensitive than standard tests if a trend in the data is lacking, (2) trends may not be linear.
5. We recommend that trend-testing be included more routinely in trial-protocols in order to increase the sensitivity of data analysis of cardiovascular trials.

6. REFERENCES

1. Kirkwood BR, Sterne JAC. Dose response relationships (trends). In: Medical statistics. Blackwell Science Malden . MA, 2003, pp 336-8.
2. SPSS Statistical Software. http://www.spss.com
3. Microsoft's Excel. www.microsoft.com
4. SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas

CHAPTER 45

ODDS RATIOS AND MULTIPLE REGRESSION MODELS, WHY AND HOW TO USE THEM

1. INTRODUCTION

In observational studies odds ratios (ORs) and multiple regressions models are commonly used for respectively the surrogate measurements of relative risks and the assessments of independent risk factors. In clinical trials both of them can be used for different purposes. Odds ratios unlike chi-square tests provide a direct insight in the strength of the relationship: odds ratios describe the probability that patients with a certain treatment will have the event compared to those without. Multiple regression models can reduce the data spread due to certain patient characteristics like differences in baseline values, and thus, improve the precision of the treatment comparison. Despite these advantages these methods are not routinely used for the evaluation of clinical trials. The current paper was written (1) to emphasize the great potential of odds ratios and multiple regression models in clinical trials, (2) to illustrate the ease of use, and (3) to familiarize the non-mathematical readership of this book with these important methods for clinical trials.

2. UNDERSTANDING ODDS RATIOS (ORS)

As stated recently by Guyatt and Rennie, while clinicians have an intuitive understanding of risks and even risk ratios, and gamblers of odds, no one, with the possible exception of a few statisticians, intuitively understands ORs.[1] The clinical perception of ORs may be difficult. Yet, they have obtained an important place in observational research, particularly, unmatched case-control studies. Because, in such studies, patients are selected on the basis of their disease, and controls are just a small sample from the target population, it is impossible to calculate either the absolute or the relative risk of a disease. Instead,

$$\text{the OR} = \frac{\text{the odds of a disease in a group exposed to a risk factor}}{\text{the odds of the same disease in a group unexposed to the risk factor}}$$

can be used as a surrogate measure for the relative risk of disease. ORs can, however, also be used for different purposes. In clinical trials, particularly, those using events as endpoints like cardiovascular trials, ORs can be used as an alternative to the traditional χ^2 – test for assessing patients with versus without an event. Apart from the p-values, χ^2 – tests do not provide an insight in the strength

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 511–527.

of the relationship. Instead, ORs measure the magnitude of association, and, in addition, describe the probability that people with a certain treatment will have the event compared to people without the treatment. Despite this advantage ORs are not routinely used for the evaluation of clinical trials.

Odds ratios(ORs) as an alternative method to χ^2 – tests for the analysis of binary data

The odds is the probability that an event happens divided by the chance that it does not so.

	event yes	no (numbers of patients)
treatment-1	p	q
treatment-2	r	s

With treatment-1 the probability or risk of an event can be described by p / (p+q) , with treatment-2 by r / (r+s), the ratio of p/(p+q) and r/(r+s) = risk ratio (RR).
The odds of an event from treatment-1 is different. It equals p/q, and the ratio of two odds, p/q and r/s is called the odds ratio (OR). In case-control studies ORs are used as a surrogate measure for RRs, because p/(p+q) in such studies is, simply, nonsense. Let us assume:

	event-group	no-event-group	target population (numbers of patients)
risk factor(treatment)	32 (p)	4(q)	4000
no risk factor(no treatment)	24 (r) +	52 (s) +	52000
	56	56	

The risk factor could be a treatment. The no-event-group group is just a random sample from the target population, but the ratio r/s is that of target population. Suppose 4 = 4000 and 52 = 52000, then $\frac{p/(p+q)}{r/(r+s)}$ is suddenly close to $\frac{p/q}{r/s}$.
This means that the OR in this situation is a good approximation of the RR of the target population. In clinical trials things are different. Both ORs and RRs can be meaningfully used. An OR or RR of 1.0 indicates no difference between treatment-1 and -2. The more distant from 1.0, the larger the difference between the two treatments where the OR is always more distant than the RR. An advantage of the RR is that it truly reflects the magnitude of the increased risk, e.g., a risk of ½ in group-1 and ¼ in group-2 produces a RR of ½ / ¼ = 2, a twice increased risk. The OR of these data produces the result 1 / (1/3) = 3, a three times increased odds ratio, which is clinically somewhat more difficult to understand. However an increased OR can still be interpreted as an increased probability of events in patients with the treatment compared to those without the treatment. For clinical trials advantages of the ORs compared to the RRs include:

1. ORs can be used as an alternative to the widely used χ^2 – tests for analyzing 2 x 2 contingency tables, while RRs can not because they use different cells.[2]
2. Statistical software uses rarely RRs, and mainly ORs. [3-10]
3. Computations using RRs are less sensitive than those using ORs. This is due to

ceiling problems, risks run from 0 to 1, odds from 0 to infinity.[11]

4. Unlike RRs, ORs are the basis of modern methods like meta-analyses of clinical trials[12], propensity scores for assessment of confounding[13], logistic regression for subgroup analysis[14], Cox regression for proportional hazard ratios[15] etc.

How to analyze odds ratios (ORs)

If we take many samples from a target population, the mean results of those samples usually follow a normal frequency distribution, meaning that the value in the middle will be observed most frequently and the more distant from the middle the less frequently a value will be observed. E.g., we will have only 5% chance to find a result more than 2 standard errors (SEs) (or more precisely 1.96 SEs) distant from the middle. The same is true with proportional data. Many statistical tests

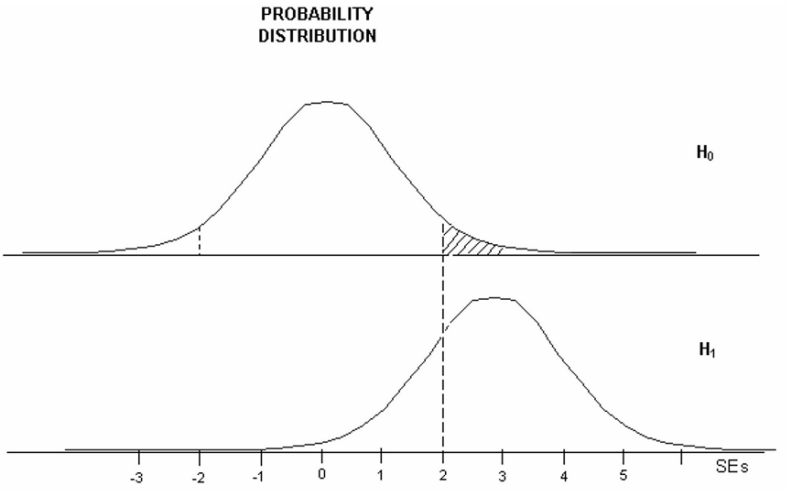


Figure 1. H1 = graph based on the data of a sample with standard errors distant from zero (SEs) as unit on the x-axis. H0 = same graph with a mean value of 0. We make a giant leap from the sample to the entire population, and we can do so because the sample is assumed to be representative for the entire population. H1 = also the summary of the means of many samples similar to our sample. H0 = also the summary of the means of many samples similar to our sample, but with an overall effect of 0. Our mean not 0 but 2.9. Still it could be an outlier of many samples with an overall effect of 0. If H0 is true, then our sample is an outlier. We can't prove, but calculate the chance/ probability of this possibility. A mean result of 2.9 SEs is far distant from 0: suppose it belongs to H0. Only 5% of H0 trials >2.0 SEs distant 0. The chance that it belongs to H0 is thus <5%. We conclude that we have <5% chance to find this result, and, therefore, reject this small chance.

make use of the normal distribution to make predictions. Figure 1 shows, e.g., how the normal distribution theorem is used to reject the null-hypothesis of no difference from zero.
A problem with ORs is that they are not normally distributed. And so, the above approach to making predictions cannot be applied. Figure 2 upper graph shows

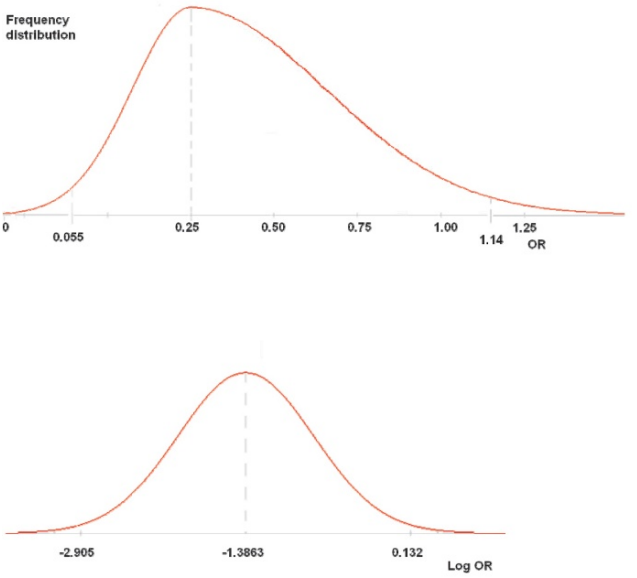


Figure 2. Upper graph: frequency distribution of an OR of 0.25 with 95% confidence interval; lower graph logarithmic transformation of the upper graph.

how skewed the frequency distribution of ORs, actually, can be. Suppose the OR of a representative sample is 0.25. Then it can be demonstrated that the chance of finding a lower or higher OR the next time are far from equal (Figure 2 upper graph). Chances can be expressed in the form of 95% confidence intervals which are for the given example between 0.055 and 1.14. With an OR of 1.803 the 95 % confidence interval is between 1.11 and 2.92 (Figure 3 upper graph). The

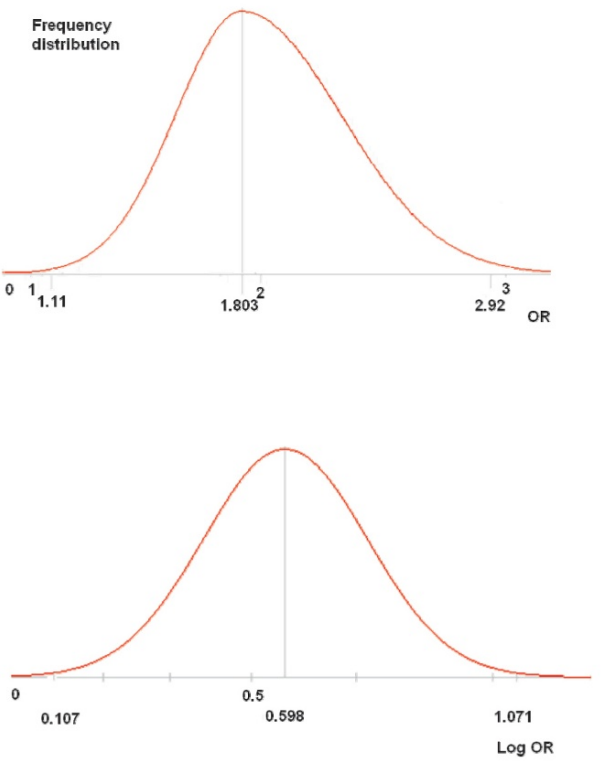


Figure 3. Upper graph: frequency distribution of an OR of 1.803 with 95% confidence interval; lower graph logarithmic transformation of the upper graph.

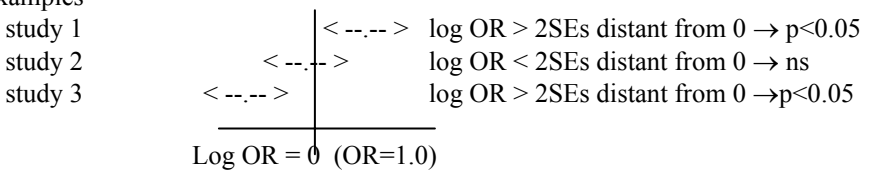
frequency distributions are not symmetrical around the observed sample OR. This asymmetry is, especially, noticeable when the sample OR is low (Figure 2 upper graph). Statisticians were very happy to observe that something wonderful happened when on the x-axis of the frequency distribution curve the OR was replaced with the logarithm of the OR (log OR). A close to normal distribution was observed (Figures 2 and 3 lower graphs). This means that the log OR can be used for testing ORs. We should add that throughout the text the term log indicates the natural logarithm (logarithm with base e).

As explained in Figure 1, if the log OR is more than 2 SEs distant from a log OR of 0, the null-hypothesis of no difference from 0 is rejected. Our result is, then, significantly different from 0 at $p < 0.05$.

	event yes	no(number of patients)
treatment-1	p	q
treatment-2	r	s

If OR(= p/q / r/s) = 1, then no difference exists between treatment-1 and -2.

If OR =1, then log OR = 0.
With normal distributions, if a mean result is > 2 SEs distant from 0, it wil be significantly different from 0 at p<0.05. Also, if log OR is > 2 SEs distant from 0, it will be significantly different from 0 at p<0.05.
Examples



In order to proceed we need to know the standard errors of the log odds ratios. For the calculation of a standard error of the log odds ratios a mathematical trick called the quadratic approximation[16] has to be used. Most functions f(x) can be represented by a power series near some point a:
$f(x) = c_0 + c_1 (x-a) + c_2 (x-a)^2 + c_3 \ldots\ldots$
where c_0 , c_1 , c_3….. are constants.
If we put x = a in the equation, then all terms after the first are 0, and $f(a) = c_0$.
If we differentiate the equation, then we have
$f'(x) = c_1 + 2 c_2 (x-a) + 3 c_3 (x-a)^2$ …..
If we put again x = a, then $f'(a) = c_1$
If we take the second differentiation, we have
$f''(x) = 2c_2 + 6 c_3 (x-a) + 12c_4 \ldots$.
$f''(a) = 2c_2$ or $c_2 = f''(a) / 2$.
As the right end terms of the equation soon will be very small, we can stop right here and neglect these terms. This means that f(x) can be described as

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2$$

Even $f(a) + f'(a)(x-a)$ is a good approximation of f (x).
This quadratic approximation formula can be conveniently used to develop a formula for the standard errors of odds ratios as follows:
$\log(x) = \log(a) + (x-a) \log'(a)$
$\log'(a)$ denotes the first derivative of log(a), the slope of the graph of log(a) against a which equals 1/a.
Adding or subtracting a constant to a variable leaves its standard error unchanged and multiplying by a constant has the effect of multiplying the standard error by that constant. Applying these rules under the assumption that the variable x is close to (a) we can further deduce:

$$se \log(x) = \frac{se\ x}{x}$$

If the variable is an odds, we can calculate the standard error of the log odds according to

$$se \log(odds) = \frac{se\ odds}{odds}$$

	number responders	non-responders
treatment-0	p	q
treatment-1	r	s

If in an experiment of (p+q) patients there are p responders to a treatment, the odds of responding is p/q. The standard error of the odds is given by the formula se odds =
$\sqrt{\ p(p+q) / q^3}$.
We can now readily calculate the standard error of the log (odds).
Se log (odds) = $\frac{\text{se odds}}{\text{odds}} = \frac{\sqrt{\ p(p+q) / q^3}}{p/q} = \sqrt{(1/p + 1/q)}$.
More relevant to us than the standard error of an odds is the standard error of an OR.
The odds of responding in treatment-group-0 is p/q, in treatment-group-1 it is r/s. The standard error of the log (OR) is given by the formula $\sqrt{(1/p + 1/q + 1/r + 1/s)}$.

Real data examples of simple OR analyses

The first example given is from the data in Figure 2 left side.

	event yes	no(numbers of patients)
treatment-1	5 (p)	10 (q)
treatment-2	10 (r)	5 (s)

OR = p/q/r/s = 0.25,
log OR = -1.3863,
SEM log OR = $\sqrt{(1/p+1/q+1/r+1/s)}$= 0.7746,
log OR ± 2 SEMs = -1.3863 ± 1.5182,
= between– 2.905 and 0.132.
Now turn the log numbers into real numbers by the antilog button of the pocket calculator.
= between 0.055 en 1.14.
This result "crosses" 1.0, and, so, it is not significantly different from 1.0.

The second example given is from Figure 3 right side.

	Event yes	no(number of patients)
treatment-1	77	62
treatment-2	103	46

OR = 103/46/77/62 = 2.239/1.242 = 1.803,
log OR = 0.589,
SEM log OR = $\sqrt{(1/103 +1/46 +1/77 +1/62)}$= 0.245,
log OR ± 2 SEMs = 0.589± 2(0.245),
= 0.589 ± 0.482,
= between 0.107 and 1.071.

Turn the log numbers into real numbers by the antilog button of the pocket calculator.

= between 1.11 and 2.92, significantly different from 1.0.

The p-value of this difference can be calculated using the t-test.
t = log OR / SEM = 0.589/0.245 = 2.4082, which according to the t-table means a p-value < 0.02.

Real data examples of advanced OR analyses

Odds ratios are also the basis of many modern methods like various logistic regression models used to adjust for subgroup analyses. A simple example of a logistic model is given.

	responders	non-responders
new treatment (group-1)	17 (p)	4 (q)
control treatment (group-2)	19 (r)	28 (s)

The odds of responding are p/q and r/s,

$$\text{odds ratio (OR)} = (p/q) / (r/s) = \frac{\text{odds of responding group-1}}{\text{odds of responding group-2}}$$

As there is a linear relationship between treatment modality and log odds of responding, we use a loglinear regression model called binary logistic regression instead of a linear regression model.
The linear regression model $y = a + bx$
is transformed into: $\log \text{odds} = a + bx$.
Log odds is the dependent variable, and x is the independent variable (treatment modality: 1 if the patient is given the new treatment, 0 if control).
Instead of $\log \text{odds} = a + bx$
We can describe the equation as $\text{odds} = e^{a+bx}$,
if new treatment, then x = 1, and $\text{odds} = e^{a+b}$,
if control treatment, then x = 0, and $\text{odds} = e^{a}$
the ratio of two treatments $\text{odds ratio} = e^{a+b} / e^{a} = e^{b}$.
Software calculates the best b for given data,
if b = 0, then $e^{b} = \text{OR} = 1$,
if b significantly > 0, then the OR significantly > 1, and there is a significant difference between the new treatment and control.
The results are

	coefficients	SEM	t	p
a	–1.95	0.53	…..	…..
b	1.83	0.63	2.9..	0.004

We can conclude that b = significantly different from 0, and that there is, thus, a significant difference between new treatment and control, the odds of cure is $e^{1.83}$ = 6.2339 times greater in the treatment group than it is in the control group.
The logistic model can adjust for subgroups as demonstrated underneath:

	responders	non-responders	responders	non-responders
	> 50 years		<50 years	
group-1 (new treatment)	4	2	13	2
group-2 (control treatment)	9	16	10	12

Software calculates best fit b- and a-values for data:

	coefficients	SEM	t	p
a>50	–2.37	0.65		
a<50	-1.54	0.59		
b>50	1.83	0.67	2.7..	0.007
b<50	1.83	0.67	2.7..	0.007

We can conclude here that the b-values are identical and both significantly different from 0. There is, thus, a significant difference between the new and control treatment also after age-class adjustment. In both subgroups the new treatment is better than control, which strengthens the earlier conclusions from these data.

3. MULTIPLE REGRESSION MODELS TO REDUCE THE SPREAD IN THE DATA

Small precision results in lack of power to reject null hypotheses and wide confidence intervals for parameter estimates. Certain patient characteristics in randomized controlled trials may cause spread in the data even if the characteristics are equally distributed among the treatment groups and do not interact with the treatment modalities. As an example, sulfonurea-compounds are efficacious for the treatment of diabetes type II. In a parallel-group clinical trial 36 patients with diabetes type II were treated with a potent (glibencamide) and a non-potent sulfonurea-compound (tolbutamide). Efficacy of treatment was assessed by fasting glucose. In the glibencamide group fasting glucose after treatment was 7.50 with standard deviation 2.01 mmol/l, in the tolbutamide group 8.50 with standard error 1.76 mmol/l. The difference in efficacy equals 8.50-7.50 = 1.00 mmol/l glucose with a pooled standard error of 0.94 mmol/l. According to the unpaired t-test this difference was not significant ($p > 0.05$). These data can also be assessed by a linear regression model with on the x-axis the treatment modality (0 = glibencamide; 1 = tolbutamide), and on the y-axis treatment efficacy (fasting glucose), (Figure 4, left graph). The regression coefficient (direction coefficient) of the regression line equals b = 1.00 mmol/l with standard error 0.94 mmol/l, p >

0.05: exactly the same result as that obtained by an unpaired t-test. However, the regression model enables to add a second variable: the presence of beta-cell failure defined as a fasting glucose >8.0 mmol/l. After adjustment for this second variable the treatment efficacy b was unchanged (1.00 mmol/l), but standard errors fell from 0.94 to 0.53 mmol/l, with a significance of difference between the two treatment modalities at $p < 0.05$ (Figure 4, right graph). This initial lack of precision was not caused by confounding, because in either subgroup the number

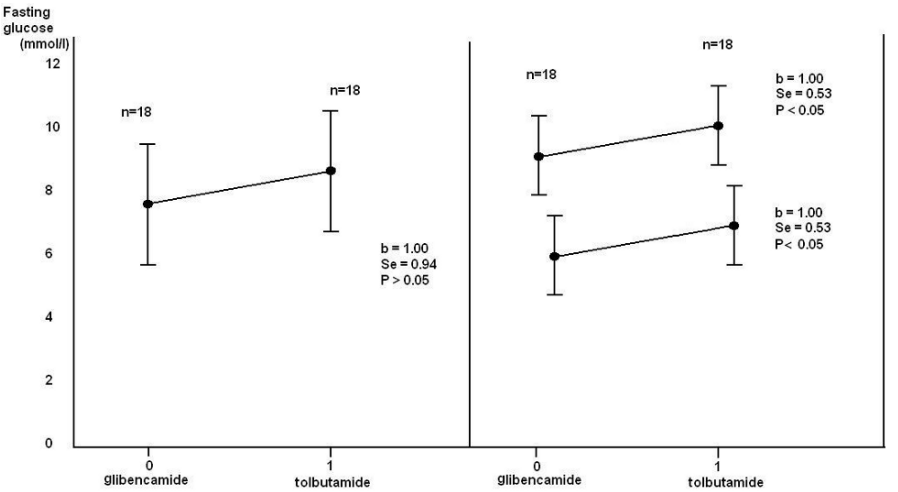


Figure 4. Mean fasting glucose levels and standard deviations of a parallel-group study of two treatments for diabetes type II. Left graph: linear regression of overall data. Right graph: the same analysis after adjustment for the presence of beta cell failure or not (fasting glucose > 8 mmol/l).

of patients receiving glibencamide was similar to that receiving tolbutamide. Also interaction could not explain the lack of precision, because the difference in treatment efficacy in the two subgroups was similar. By appropriate data modeling some of the variability is removed from the data, and a more precise data comparison is produced. So far, data modeling has not been emphasized in the analysis of prospective randomized clinical trials, and special statistical techniques need to be applied including the transformation of parallel-group data into regression data and the addition of covariates to the models.[15,17] We should emphasize that low precision in a clinical trial may be caused by a biased study due to the presence of confounding or interacting variables. Adjusting such variables will, of course, improve precision. In the present study we demonstrate that some

covariates even if they are no confounding or interacting variables, may contribute to increasing precision of the data analysis.

A linear regression model for increasing precision

The underneath data present a parallel-group trial comparing efficacy of a new laxative versus control laxative.

patient no.	treatment modality (new=0, control=1)	response = stool frequency after treatment (4 week stools)	baseline stool frequency (4 week stools)
1	0	24	8
2	0	30	13
3	0	25	15
4	1	35	10
5	1	39	9
6	0	30	10
7	0	27	8
8	0	14	5
9	1	39	13
10	1	42	15
11	1	41	11
12	1	38	11
13	1	39	12
14	1	37	10
15	1	47	18
16	0	30	13
17	1	36	12
18	0	12	4
19	0	26	10
20	1	20	8
21	0	43	16
22	0	31	15
23	1	40	14
24	0	31	7
25	1	36	12
26	0	21	6
27	0	44	19
28	1	11	5
29	0	27	8
30	0	24	9
31	1	40	15
32	1	32	7
33	0	10	6
34	1	37	14
35	0	19	7

SPSS statistical software is used for analysis.[10]
First, enter the data or a data-file, e.g., from Excel.[18]
Then command: statistics; regression; linear.

The underneath results are presented.
The mean difference in response between new treatment and control = 9.824 stools per 4 weeks (Se = 2.965). The t-test produces a t-value of 9.824 / 2.965 = 3.313, and the t-table gives a p-value of <0.01.
A linear regression according to

$y = a + bx$
with y = response and x = treatment modalities (0 = new treatment, 1 = control),
a = intercept, and b= regression coefficient,

produces a similar result

$b = 9.824$
$se_b = 2.965$
$t = 3.313$
p-value = 0.020
95 % confidence interval 4.013-15.635.

Improved precision of this data analysis is a possibility if we extend the regression model by including a second explanatory-variable = baseline stool frequency according to

$y = a + b_1 x_1 + b_2 x_2$
with x_1 = treatment modalities (0 = new treatment, 1 = control),
x_2 = baseline stool frequencies, and b-values are partial regression coefficients.

This produces the following results

$b_1 = 6.852$
$se_{b1} = 1.792$
$t = 3.823$
p-value =0.001
95 % confidence intervals 3.340-10.364.

Now, the 95 % confidence interval for the treatment effect is substantially narrower than the previously presented confidence interval. So, by adjusting for the baseline stool frequencies an improved precision is obtained as demonstrated by a

smaller confidence interval, a larger t-value, and a smaller p-value. We should of course answer the questions: is baseline stool a (1) confounding or (2) interacting variable. For answering question (1) we perform a simple linear regression analysis of the variables x_1 versus x_2 which shows that the two variables are independent of one another ($P>0.05$). X_2 is, thus, not a confounding variable. For answering question (2) a multiple linear regression is used with x_1, x_2 and x_3 as interacting variable given by $x_1 . x_2$ (x_1 times x_2). This analysis shows that x_3 is not a significant determinant of treatment response ($p>0.05$). There is, thus, no interaction between the two independent variables in the model. This means that increased precision to predict treatment response is obtained by including the baseline stool into the model, and that this model is otherwise unbiased by confounding or interaction.

A logistic regression model for increasing precision

Consider the underneath two by two contingency table.

	Numbers Responders	numbers non-responders
Treatment 1	30 a	45 b
Treatment 2	45 c	30 d

The odds-ratio-of-responding equals a/b / c/d = 30/45 / 45/30 = 0.444. The natural logarithmic (ln) transformation of this odds ratio equals -0.8110. The approximate standard error of this logarithmic transformation is given by
$\sqrt{}$ (1/a +1/b+1/c+1/d) = $\sqrt{}$ (1/30+1/45+/1/30+1/45)= 0.333.
A t-test of these data produces a t-value of 0.8110/0.333= 2.435. According to the t-table this odds-ratio is significantly different from an odds ratio of 1.0 with a p-value of 0.015.
Logistic regression according to the model

ln odds-of-responding = a + bx
with x = treatment modality (0 or 1),
a = intercept, and b = regression coefficient,

produces the same result.

SPSS statistical software is again used to calculate the best b-values for the data given.
First enter the data or an Excel data file.
The command: statistics; regression; binary logistic.
The underneath results are presented.

b = 0.8110
se_b = 0.333

odds ratio of responding with treatment 1 / treatment 2 = 2.250
with 95% confidence interval 1.613-3.139
p-value = 0.015

The patients can be divided into two age classes:

	Over 50 years		Under 50 years	
	Responders	non-responders	responders	non-responders
Treatment 1	18	20	12	25
Treatment 2	31	8	14	22

Improved precision of the statistical analysis is a possibility if we control for age groups using the underneath multiple logistic regression model

ln odds-of-responding = $a + b_1 x_1 + b_2 x_2$
with x_1 = treatment modalities (0= treatment 1, 1= treatment 2)
x_2 = age classes (0= < 50 years, 1 = > 50 years)
b-values are regression coefficients.

The following results are obtained:

b_1 =0.867
se_{b1} =0.350
odds ratio of responding with treatment 1 / treatment 2 = 2.380
with 95% confidence interval 1.677-3.377
p-value = 0.012

After adjustment for age class improved precision to test the efficacy of treatment has been obtained as demonstrated by a smaller p-value. Is this increased precision due to unmasked confounding or interaction? For answering these questions we perform a simple binary logistic regression of the variables x_1 versus x_2 which shows that the two variables are independent of one another (P>0.05). X_2 is not a confounding variable. A multiple binary logistic regression is used with x_1 , x_2 and x_3 as interacting variable given by x_1 . x_2 (x_1 times x_2).This analysis shows that x_3 is not a significant determinant of treatment response (p>0.05). There is, thus, no interaction between the two independent variables in the model. This means increased precision to predict treatment response has been obtained by including the age-category as covariate into the model, and that, like the previous example, it is unbiased by confounding or interaction.

4. DISCUSSION

Advantages of the ORs compared to the RRs include (1) that, unlike RRs, they can be used as an alternative to the widely used χ^2 – tests for analyzing binary data in clinical trials, (2) that software for ORs is widely available, (3) that unlike RRs, ORs do not suffer from ceiling problems, and (4) that they are the basis of many modern methods like logistic regression, and Cox regression. An advantage of ORs compared to the traditional χ^2 – tests is that ORs provide, in addition to p-values, a direct insight in the strength of the relationship: odds ratios describe the probability that people with a certain treatment will have the event compared to people without the treatment.

For the analysis of ORs the logarithms of the ORs should be used. Data results are obtained by turning the logarithmic numbers into real numbers by using their antilogarithms.

A limitation of the ORs is that, although they adequately present the relative benefits of a treatment compared to control, they do not tell us anything about the absolute benefits. For that purpose information about baseline risks or likelihoods are required. E.g., with an odds ratio of cure of treated versus baseline of 5, and a baseline likelihood of cure of 10 out of 1000 patients, the number of cured will increase to approximately 50 out of 1000, with a baseline of 100 out of 1000 patients, it will do so to approximately 500 out of 1000.

ORs, despite a fairly complex mathematical background, are easy to use, even for non-mathematicians, and they are the basis of many modern methods for analyzing clinical data including multivariable methods.

Multiple regression analysis of confounding variables, although routinely used in retrospective observational studies, is not emphasized in prospective studies like randomized clinical trials (RCTs). The randomization process ensures that differences in potential confounders are the result of chance. If differences are statistically significant, multiple regression analysis can be used for adjustment. Multiple regression can, also, be used in prospective studies for a different purpose. Certain patient characteristics in RCTs may cause substantial spread in the data even if they are equally distributed. Including such data in the efficacy analysis may increase precision and power in the data analysis. When the dependent variable is a change score, as in the first example, the baseline level is the first candidate to be considered, because it is almost certainly associated with the change score. When the dependent variable is an odds ratio, like in the second example, gender or age-category are adequate candidates.

We should emphasize that it has to be decided prior to the trial and stated explicitly in the trial protocol whether a regression model will be applied, because post hoc decisions regarding regression modeling like any other post hoc change in the raises the risk of statistical bias due to multiple testing. Naturally, there is less opportunity for modeling in a small trial than in a large trial. There is no general rule about which sample sizes are required for sensible regression modeling, but one rule-of-thumb is that at least ten times as many patients are required as the number of variables in the model. This would mean that a data set of at least 30 is

required if we wish to include a single covariate in the model for the purpose of improving precision. With every additional covariate in the model an extra regression weight must be estimated, which rapidly leads to a decreased rather than improved precision.
Regression analysis can be adequately used for improving precision of efficacy analysis. Application of these models is very easy since many computer programs are available. For a successful application the fit of the regression models should, however, always be checked for example by scatter plots, or in case of doubt by goodness of fits tests, and the covariate selection should be sparse.
We do hope that this paper will stimulate clinical investigators to use odds ratios and multiple regression models more often.

5. CONCLUSIONS

Odds ratios (ORs) unlike χ^2 – tests provide a direct insight in the strength of the relationship between treatment modalities and treatment effects. Multiple regression models can reduce the data spread due to certain patient characteristics, and thus, improve the precision of the treatment comparison. Despite these advantages the use of these methods in clinical trials is relatively uncommon.
This chapter (1) emphasizes the great potential of odds ratios and multiple regression models as a basis of modern methods, (2) illustrates their ease of use, and (3) familiarizes the non-mathematical scientific community with these important methods.
Advantages of the ORs are multiple:
1. They describe the probability that people with a certain treatment will have an event compared to people without the treatment, and are, therefore, a welcome alternative to the widely used χ^2 – tests for analyzing binary data in clinical trials.
2. Statistical software of ORs is widely available.
3. Computations using risk ratios (RRs) are less sensitive than those using ORs.
4. ORs are the basis for modern methods like meta-analyses, propensity scores, logistic regression, Cox regression etc.
For analysis logarithms of the ORs have to be used, results are obtained by calculating antilogarithms. A limitation of the ORs is that they present relative benefits but not absolute benefits. ORs, despite a fairly complex mathematical background, are easy to use, even for non-mathematicians.
Both linear and logistic regression models can be adequately applied for the purpose of improving precision of parameter estimates like treatment effects. We caution that, although application of these models is very easy with computer programs widely available, the fit of the regression models should always be carefully checked, and the covariate selection should be carefully considered and sparse.
We do hope that this paper will stimulate clinical investigators to use odds ratios and multiple regression models more often.

6. REFERENCES

1. Guyatt G, Rennie D. Users guide to the medical literature-a manual for evidence based clinical practice by the Evidence-Based Medicine Working Group Cicago, USA. AMA Press. 2001; pp 356-7.
2. Bland JM, Altman DG. The odds ratio. BMJ 2000; 320: 1468.
3. BUGS y WinBUGS. http://www.mrc-bsu.cam.ac.uk/bugs http://cran.r -project.org
4. S-plus.http://www.mathsoft.com/splus
5. Stata. http://www.stata.com
6. StatsDirect. http://www.camcode.com
7. StatXact. http://www.cytel.com/products/statxact/statact1.html
8. True Epistat. http://ic.net/~biomware/biohp2te.htm
9. SAS. http://www.prw.le.ac.uk/epidemiol/personal/ajs22/meta/macros.sas
10. SPSS Statistical Software. http://www.spss.com
11. Zwinderman AH, Niemeijer MG, Kleinjans HA, Cleophas TJ. Application of item response modelling for quality of life assessments: effect of two nitrate treatment regimens in stable angina pectoris. In: What should a clinical pharmacologist know to start a clinical trial, Kuhlmann and Mrozikiewicz, editors. Zuckschwerdt Verlag, New York, 1998, pp 40-8.
12. Zwinderman AH. Subgroup analysis using multiple regression, confounding, interactions, synergism. In: Statistics applied to clinical trials. Edited by Cleophas, Zwinderman and Cleophas, Springer 2006; Dordrecht, Neth, pp 125 -40.
13. Cleophas TJ, Zwinderman AH. Statistical primer for cardiovascular research, meta-analysis. Circulation 2007; 115: 2870-5.
14. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997; 127: 757-63.
15. Cleophas TJ. Problems with regression modeling for the analysis of clinical trials. Int J Clin Pharmacol Ther 2005; 43: 23-9.
16. Anonymous. Explanation of the quadratic approximation. www.math.tamu.edu/~kfu/Quadratic Approximation.pdf
17. Cleophas TJ. The sense and non-sense of regression modeling for increasing precision of clinical trials. Clin Pharmacol Ther 2003; 74: 295-7.
18. Excel by Windows. http://www.excel.com

CHAPTER 46

STATISTICS IS NO "BLOODLESS" ALGEBRA

1. INTRODUCTION

Because biological processes are full of variations, statistics can not give you certainties, but only chances. What kind of chances? Basically, the chances that prior hypotheses are true or untrue. The human brain excels in making hypotheses. We make hypotheses all the time, but they may be untrue. E.g., when you were a kid, you thought that only girls could become doctors, because your family doctor was a girl. Later on this hypothesis appeared to be untrue. In clinical medicine we currently emphasize that hypotheses may be equally untrue and must be assessed prospectively with hard data. That's where statistics comes in, and that is where at the same time many a clinician starts to become nervous, loses his / her self-confidence, and is more than willing to leave his / her data to the statistical consultant who subsequently runs the data through a whole series of statistical tests of SAS[1] or SPSS[2] or comparable statistical computer software to see if there are any significances. The current article was written to emphasize that the above scenario of analyzing clinical trial data is bad practice and frequently kills the data, and that biostatistics can do more for you than provide you with a host of irrelevant p-values.

2. STATISTICS IS FUN BECAUSE IT PROVES YOUR HYPOTHESIS WAS RIGHT

Statistics is fun, particularly, for the clinical investigator. It is not mathematics, but a discipline at the interface of biology and mathematics. This means that maths is used to answer the biological questions. The scenario as described above does not answer reasonable biological questions. It is called data dredging and is the source of a lot of misinterpretations in clinical medicine. A statistical analysis should be confined to testing of the prior hypotheses. The problem with multiple statistical tests can be explained by gambling 20 times with a chance of success of 5%. You can be sure that after the game you will get $(1\text{-}0.05)^{20} = (0.95)^{20} = 0.36 = 36\%$ chance to win a prize. This result is, however, not based on any significant effect but rather on the play of chance. Now, don't let it happen to your trial. Also, a statistical result that does not confirm your prior belief, don't trust it. Make sure that the simplest univariate tests are used for your prospective trial data, because they are adequate and provide the best power. Fancy multivariate procedures are not in place to answer your prior hypotheses. Statistics is fun, because it generally confirms or

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 529–535.

largely confirms your prior hypotheses, which is appropriate because they were based on sound clinical arguments. If they don't, this is peculiar and should make you anxious to find out why so: imperfections within the design or execution of the trial?[3] It is fun to prove your hypothesis was right, or to find out what you did overlook. Another fun thing with statistics, although completely different and by far not so important, is the method of secondary analyses: it does not prove anything, but it is kind of sports and gives you new and sound ideas for further research.

3. STATISTICAL PRINCIPLES CAN HELP TO IMPROVE THE QUALITY OF THE TRIAL

Over the past decades, the randomized controlled trial has entered an era of continuous improvement and has gradually become accepted as the most effective way of determining the relative efficacy and toxicity of a new therapy, because it controls for placebo and time effects. However, even sensitive and properly designed and executed trials do not always confirm hypotheses to be tested, and conclusions are not always confirmed by subsequent trials. Although the former may be due to wrong hypotheses, the latter is likely to be due to the presence of certain imperfections within the design and execution, and analysis of the trial itself. Such principles could include[4] : (1) giving every effort to avoid asymmetries in the treatment groups (chapter 1, stratification issues), (2) emphasis on statistical power rather than just null-hypothesis testing (chapter 5), (3) assessing asymmetries of outcome variables in order to determine the most important determinants of clinical benefit (chapter 16), (4) accounting routinely for Type III errors of mistakenly believing that an inferior treatment is superior (chapter 5), (5) routinely weighing the odds of benefits against the odds of risks of new treatments.

4. STATISTICS CAN PROVIDE WORTHWHILE EXTRAS TO YOUR RESEARCH

The classical two-parallel-groups design for clinical drug trials is a rather dull activity and is, essentially, unable to answer many current scientific questions. Also, it is laborious, and in the clinical setting sometimes ethically or financially impossible. Examples of what the classical clinical trial design cannot manage: (1) assess multimodal therapies, (2) account historical data, (3) safeguard ethics and efficacy during the course of long-term trials, (4) study drugs, before well-established toxicity information is available, (5) account the possibility of therapeutic equivalence between test and reference treatment, (6) study multiple treatments in one trial, (7) adjust change scores for baseline levels. Alternative designs for such purposes: (1) factorial designs (chapter 1)[5], (2) historical controls designs (chapter 1)[6], (3) group-sequential interim analysis designs (chapter 6)[7], (4) sequential designs for continuous monitoring (chapter 6)[8], (5) therapeutic equivalence designs (chapter 4), (6) multiple crossover-periods / multiple parallel-groups designs[9], (7) increased precision designs through multivariate adjustment

(chapter 12). There is, of course, the increased risks of type I/II errors, and the possible loss of some of the validity criteria with the novel designs. However, provided that such possibilities are adequately accounted for in the design stage of the trial, the novel designs are acceptedly valid, and offer relevant scientific, ethical, and financial extras.

5. STATISTICS IS NOT LIKE ALGEBRA BLOODLESS

Statistics is not like algebra bloodless, and requires a lot of biological thinking and just a little bit of mathematics. For example, mathematically we need representative sample sizes to make meaningful inferences about the whole population. Yet, from a biological point of view, this is less true: the first datum encountered in a clinical situation of complete ignorance provides the greatest amount of information from any one datum an investigator will encounter. E.g., consider a new disease for which there is no knowledge whatsoever about the order of magnitude of time of exposure, time of incubation, time of appearance of subsequent symptoms. The first patient for whom we know such data provides a great deal of information.

Another example of biological rather than mathematical thinking involves the issue of making the test parameters alpha and beta flexible. They are mostly set at respectively 5 and 20%. A 20% beta is, however, larger than is appropriate in many cases. E.g., when the false positive is worse for the patient than the false negative, as in case of testing a drug for non-life threatening illness with the drug having severe side effects, the 5 and 20% choices for alpha and beta are reasonable. However, in testing treatment for cancer, the rate of false negatives is worse for the patient, and so, the ratio beta/alpha should be reduced.

A third example of biological thinking is the inclusion of a "safety factor" when estimating prior to a trial the sample size required. Usually the required sample size is calculated from a pilot study or from results quoted in the literature. However, these data are not the actual data from our study, and not using the real data may introduce an error. Also, the data as used for sample size calculation are subject to randomness error. Due to such errors the alpha and beta errors upon which our sample size is based may be larger than we thought. Because of these possibilities we should add a "safety factor" to the sample size as calculated, and make our sample size somewhat larger than the calculated one, e.g., 10 % larger. This is more important, the more uneasy we are about the ultimate result of the study being in agreement with the estimate used for sample size calculation.

6. STATISTICS CAN TURN ART INTO SCIENCE

Traditionally, the science of medicine is considered to be based on experimental evidence, while the art of science is supposed to be based on trust, sympathy, the threatened patient, and other things that no one would believe that could ever be estimated by statistical methods. It is true that factors, of psychosocial and personal nature, are difficult to measure, but it is not impossible to do so. At first, quality of life assessments were based on the amount of primary symptoms, e.g., pain scores etc. Increasingly it is recognized that it should be based on factors like feeling of well-being, social performance. Along this line of development, the art of medicine is more and more turned into science, e.g., with modern quality of life assessments addressing general feeling of well-being, physical activity domains etc. Statistical analyses can be readily performed on validated quality of life indices or any other measurements of effectiveness as developed [chapter 15]. It follows that this development is going to accomplish something that was only shortly believed to be impossible: turning the art of medicine into the science of medicine.

7. STATISTICS FOR SUPPORT RATHER THAN ILLUMINATION?

In 1948 the first randomized controlled trial was published.[10] Until then, observations had been largely uncontrolled. Initially, trials frequently did not confirm hypotheses to be tested. This phenomenon was attributed to little sensitivity due to small samples, as well as inappropriate hypotheses based on biased prior trials. Additional flaws were being recognized and, subsequently better accounted for: carryover effects due to insufficient washout from previous treatments, time effects due to external factors and the natural history of conditions being studied, bias due to asymmetry between treatment groups, lack of sensitivity due to a negative correlation between treatment responses etc. Currently, due to the complexity of trials, clinicians increasingly leave the thinking to statisticians, a practice which is essentially wrong and produces flawed research, because bio-research requires a lot of biological thinking and no more than a bit of statistics. Moreover, a statistician can do much more for you than provide you with a host of irrelevant p-values, but he/she can only do so, if you intuitively know what statistics can and what it cannot answer. Like Professor M. Hills, the famous statistician of London, used to say, clinicians often use statistics as a drunk uses a lantern standard, for support rather than illumination. Illumination can be obtained by exploring your clinical intuition against a mathematical background.

8. STATISTICS CAN HELP THE CLINICIAN TO BETTER UNDERSTAND LIMITATIONS AND BENEFITS OF CURRENT RESEARCH

Medical literature is currently snowed under with mortality trials, showing invariably a highly significant 10-30% relative increase in survival. Mortality is considered an important endpoint, and this may be so. Yet, a relative increase in survival of 10-30% generally means in absolute terms an increase of no more than 1-2%. Mortality is an insensitive variable of the effects of preventive medicine that is begun when subjects are middle-aged. At such ages the background noise associated with senescence becomes high. The endpoints better be reduction in morbidity so far. In addition, many clinicians know that patients would prefer assessment of quality of life and reduced morbidity rather than 1-2% increased survival in return for long term drug treatment with considerable side effects. Relative risk reductions are frequently overinterpreted by clinicians in terms of absolute risk reductions. And so are underpowered p-values: a p-value of 0.05 after all means the chance of a type II error of 50%.

On the other hand, statistics can do a lot more for clinicians than calculating p-values and relative risk reductions. Multivariate analyses can be used not only for exploring new ideas, but also for increasing precision of point estimates in a trial. Benefit/risk analyses of trial data are helpful to provide relevant arguments for clinical decision making, and they are particularly so when their ratios is assessed quantitatively. Statistics can provide us with wonderful meta-analyses of independent trials to find out whether scientific findings are consistent and can be generalized across populations.

9. LIMITATIONS OF STATISTICS

Of course, we should avoid giving a non-stop laudatio of statistics only. It is time that we added a few remarks on its limitations and possible disadvantages in order to express a more balanced opinion. Statistics is at the interface of mathematics and biology. Therefore, it gives no certainties, only chances. What chances? E.g., chances that hypotheses are true or untrue. We generally reject the null-hypothesis of no effect at $p < 0.05$. However, $p = 0.05$ means 5% chance of a type I error of finding a difference where there is none, and 50% chance of a type II error of finding no difference where there is one. It pictures pretty well how limited statistical inferences can be. In addition to the risks of type I and type II errors, there is the issue of little clinical relevance in spite of statistically significant findings. A subanalysis of the SOLVD study[11] found no symptoms of angina pectoris in 85.3% of the patients on enalapril and in 82.5% of the patients on placebo (difference statistically significant at $p < 0.001$). In situations like this, one has to wonder about the clinical relevance of the small difference. This is even more so when one considers that an active compound generally causes more side-effects than does a placebo. Finally, we have to consider the point of bias. Arguments have been raised that controlled clinical trials although they adjust for placebo effects and time effects, are still quite vulnerable to other biases, e.g., psychological biases

and selection biases. In clinical trials, as opposed to regular patient care, patients are generally highly compliant; their high compliance is an important reason for participating in the trials in the first place. They have a positive attitude towards the trial and anticipate personal benefit from it, a mechanism which is known as the Hawthorne effect.[12] Alternatively, patients selected for a trial often refuse to comply with randomization which may render unrepresentative samples.[13] Statistics has great difficulty in handling such effects and is, essentially, unable to make sense of unrepresentative samples. Not being familiar with statistics raises a two-way risk: you're not only missing the benefit of it but also fail to adequately recognize the limitations of it.

10. CONCLUSIONS

1. Statistics is fun for the clinical investigator because it generally confirms or largely confirms his / her prior hypotheses.
2. Accounting some simple statistical principles can help the clinical investigator reduce imperfections in the design and execution of clinical trials.
3. For the clinical investigator getting a good command of non-classical study designs can provide worthwhile extras to his / her research.
4. Statistics is not like algebra, because it requires a lot of biological thinking and just a little bit of mathematics.
5. Statistical analyses can be readily performed on such modern quality of life assessments like general feeling of well-being, physical activity domains, psychosocial performance etc.
6. Along this line the art of medicine is more and more being turned into scientific evidence.
7. Statistics can do a lot for the clinical investigator if he / she intuitively knows what statistics can and what it cannot answer.
8. Statistics can help clinical investigators to interpret more adequately limitations as well as benefits of current clinical research.
9. Statistics has, of course, limitations of its own. It can not give certainties, only chances.
10. Statistical significance does not automatically indicate clinical relevance.
    Statistical methods can not test every possible source of bias in a trial.

Not being familiar with statistics raises a two-way risk: you're not only missing the benefit of it but also fail to adequately recognize its limitations. We hope that this book will be an incentive for readers to improve their statistical skills in order to better understand the statistical data as published in the literature and to be able to take better care of their own experimental data.

11. REFERENCES

1. SAS Statistical Software 2000 New York, NY, USA.
2. SPSS Statistical Software 2000 Chicago, IL, USA.
3. Cleophas TJ. Methods for improving drug trials. Clin Chem Lab Med 1999; 37: 1035-41.
4. Cleophas TJ, Zwinderman AH. Limits of randomized trials, proposed alternative designs. Clin Chem Lab Med 200; 38: 1217-23.
5. Farewell VT, Dángio GJ. Report of the National Wilms´ Tumor Study Group. Biometrics 1981; 37: 169-76.
6. Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. Am J Med 1982; 72: 233-40.
7. Pocock SJ 1988 Clinical trials. A practical approach. Wiley, New York, USA
8. Whitehead J 1998 Planning and Evaluating Sequential Trials (PEST, version 3). Reading: University of Reading (www.reading.ac.uk/mps/pest/pest.html)
9. Lauter J. Exact t and F-tests for analyzing studies with multiple endpoints. Biometrics 1996; 52: 964-70.
10. Medical Research Council. Streptomycin Treatment of pulmonary tuberculosis. Br Med J 1948; 2: 769 –82.
11. Yusuf S, Pepine CJ, Garces C. Effect of enalapril on myocardial infarction and angina pectoris in patients with low ejection fraction. Lancet 1992; 340: 1173-8.
12. Campbell JP, Maxey VA, Watson WA. Hawthorne effect: implications for prehospital research. Ann Emergency Med 1995; 26: 590-4.
13. Cleophas TJ. The use of a placebo-control group in clinical trials. Br J Clin Pharmacol 1997; 43: 219-21.

CHAPTER 47

BIAS DUE TO CONFLICTS OF INTERESTS, SOME GUIDELINES

1. INTRODUCTION

The controlled clinical trial, the gold standard for drug development, is in jeopardy. The pharmaceutical industry rapidly expands it's commend over clinical trials. Scientific rigor requires independence and objectivity. Safeguarding such criteria is hard with industrial sponsors, benefiting from favorable results, virtually completely in control. The recent Good Clinical Practice Criteria adopted by the European Community[1] were not helpful, and even confirmed the right of the pharmaceutical industry to keep everything under control. Except for the requirement that the trial protocol should be approved by an external protocol review board, little further external monitoring of the trial is required in Europe today. The present paper was written to review flawed procedures jeopardizing the credibility of current clinical trials, and to look for possible solutions to the dilemma between sponsored industry and scientific independence.

2. THE RANDOMIZED CONTROLLED CLINICAL TRIAL AS THE GOLD STANDARD

Controlled clinical trials began in the UK with James Lind, on H.M.S. Salisbury, a royal Frigate, by the end of the 18th century. However, in 1948 the first randomized controlled trial was actually published by the English Medical Research Council in the British Medical Journal.[2] Until then, published observations had been uncontrolled. Initially, trials frequently did not confirm hypotheses to be tested. This phenomenon was attributed to little sensitivity due to small samples, as well as inappropriate hypotheses based on biased prior trials. Additional flaws were being recognized and, subsequently were better accounted for: carryover effects due to insufficient washout from previous treatments, time effects due to external factors and the natural history of the condition under study, bias due to asymmetry between treatment groups, lack of sensitivity due to a negative correlation between treatment responses etc. Such flaws mainly of a technical nature have been largely implemented and lead to trials after 1970 being of significantly better quality than before. And so, the randomized clinical trial has gradually become accepted as the most effective way of determining the relative efficacy and toxicity of new drug therapies. High quality criteria for clinical trials include clearly defined hypotheses, explicit description of methods, uniform data analysis, but, most of all, a valid

T.J. Cleophas et al., Statistics Applied to Clinical Trials, 537–543.

design. A valid design means that the trial should be made independent, objective, balanced, blinded, controlled, with objective measurements. Any research but, certainly, industrially-sponsored drug research where sponsors benefit from favorable results, benefits from valid designs.

3. NEED FOR CIRCUMSPECTION RECOGNIZED

The past decade focused, in addition to technical aspects, on the need for circumspection in planning and conducting clinical trials.[3] As a consequence, prior to approval, clinical trial protocols started to be routinely scrutinized by different circumstantial organs, including ethic committees, institutional and federal review boards, national and international scientific organizations, and monitoring committees charged with conducting interim analyses. And so things seems to be developing just fine until something else emerged, the rapidly expanding commend of the pharmaceutical industry over clinical trials. Scientific rigor requires independence and objectivity of clinical research, and safeguarding such principles is hard with sponsors virtually completely in control.

4. THE EXPANDING COMMEND OF THE PHARMACEUTICAL INDUSTRY OVER CLINICAL TRIALS

Today megatrials are being performed costing billions of dollars paid by the industry. Clinical research has become fragmented among many sites, and the control of clinical data often lies exclusively in the trial sponsor's hands.[4] A serious issue to consider here is adherence to scientific criteria like objectivity, and validity criteria like blindness during the analysis phase. In the USA, the FDA audits ongoing registered trials for scientific validity. However, even on-site-audits can hardly be considered capable of controlling each stage of the trial. Not any audits are provided by the FDA's European counterparts. Instead, in 1991, the European Community endorsed the Good Clinical Practice (GCP) criteria developed[1] as a collaborative effort of governments, industries, and the profession. For each of the contributing parties benefits are different. Governments are interested in uniform guidelines and uniform legislation. For the profession the main incentives are scientific progress, and the adherence to scientific and validity criteria. In contrast, for the pharmaceutical industry a major incentive is its commercial interest. And so, the criteria are, obviously, a compromise. Scientific criteria like clearly defined prior hypotheses, explicit description of methods, uniform data analyses are broadly stated in the guidelines given.[1] However, scientific criteria like instruments to control independence and objectivity of research are not included. Validity criteria like control groups and blinding are recognized, but requirements like specialized monitoring teams consistent of a group of external independent investigators guiding such criteria, and charged with interim analysis and stopping rules are not mentioned. And so, the implementation of the Good Clinical Practice Criteria is not

helpful for the purpose of safeguarding scientific independence. Instead, they confirmed the right of the pharmaceutical industry to keep everything under control.

5. FLAWED PROCEDURES JEOPARDIZING CURRENT CLINICAL TRIALS

Flawed procedures jeopardizing current clinical trials are listed in table 1. Industries, at least in Europe, are allowed to choose their own independent protocol review board prior to approval. Frequently, a pharmaceutical company chooses one-and-the-same-board for all of its (multicenter) studies. The independent protocol review board may approve protocols, even if the research is beyond its scope of expertise, for example, specialized protocols like oncology-protocols without an oncologist among its members. Once the protocol is approved, little further external review is required in Europe today. Due to recent European Community Regulations, health facilities hosting multicenter trials are requested to refrain from scientific or ethic assessment. Their local committees may assess local logistic aspects of the trial but no more than that. And so, the once so important role of local committees to improve the objectivity of sponsored research is minimized. Another problem with the objectivity of industrially-sponsored clinical trials is the fact that the trial monitors are often employees of the pharmaceutical industry. Furthermore, data control is predominantly in the hands of the sponsor. Interim analyses are rarely performed by independent groups. The scientific committee of the trial consists largely of prominent but otherwise uninvolved physicians attached to the study, the so-called *guests*. Analysis and report of the trial is generally produced by clinical associates at the pharmaceutical companies, the *ghosts*, and, after a brief review, co-signed by prominent physicians attached to the study the so-called *graphters*.

Table 1. Flawed procedures jeopardizing current clinical trials

1.	Pharmaceutical industries, at least in Europe, are allowed to choose their own independent review board prior to approval.
2.	the independent protocol review board approves protocol even if the research is beyond the scope of its expertise.
3.	Health facilities hosting multicenter research are requested to refrain from ethic or scientific assessment after approval by the independent review board.
4.	Trial monitors are often employees of pharmaceutical industry.
5.	Data control is predominantly in the hands of the sponsor.
6.	Interim analyses are rarely performed by independent groups.
7.	The scientific committee of a trial consists largely of guests (names of prominent physicians attached to the study) and graphters (for the purpose of giving the work more impact).
8.	The analysis and report is produced by *ghosts* (clinical associates at the pharmaceutical companies) and is after a brief review co-signed by the *guests* and *graphters*.

6. THE GOOD NEWS

The Helsinki guidelines rewritten in the year 2000 have been criticized[5] for its incompleteness regarding several ethical issues, e.g., those involving developing countries. However, these independently written guidelines also included important improvements. For the first time the issue of conflict of interests has been assessed in at least 5 paragraphs. Good news is also the American FDA's initiative to start auditing sponsored trials on site. In May 1998 editors of 70 major journals have endorsed the Consolidated Standards of Reporting Trials Statement (CONSORT) in an attempt to standardize the way trials are conducted, analyzed and reported. The same year, the Cochrane Collaborators together with the British journals The Lancet and The British Medical Journal have launched the "Unpublished Paper Amnesty Movement", in an attempt to reduce publication bias. There is also good news from the basis. E.g., in 30 hospitals in the Netherlands local ethic committees, endorsed by the Netherlands Association of Hospitals, have declared that they will not give up scrutinizing sponsored research despite approval by the independent protocol review board.

In our educational hospital house officers are particularly critical of the results of industrially-sponsored research even if it is in the Lancet or the New England Journal of Medicine, and they are more reluctant to accept results not fitting in their prior concept of pathophysiology, if the results are from industrially-sponsored research. Examples include: ACE-inhibitors for normotensive subjects at risk for cardiovascular disease (HOPE Study[6]), antihypertensive drugs for secondary secondary prevention of stroke in elderly subjects (PROGRESS Study[7]), beta-blockers for heart failure (many sponsored studies, but none of them demonstrating an unequivocal improvement of cardiac performance[8]), cholesterol-lowering treatment for patients at risk of cardiovascular disease but normal LDL-cholesterol levels (Heart Protection Study), hypoglycemic drugs for prediabetics (NAVIGATOR Study). As a matter of fact, all of the above studies are based on not so sensitive univariate analyses. When we recently performed a multivariate analysis of a secondary prevention study with statins, we could demonstrate that patients with normal LDL-cholesterol levels did not benefit.[9]

7. FURTHER SOLUTIONS TO THE DILEMMA BETWEEN SPONSORED RESEARCH AND THE INDEPENDENCE OF SCIENCE

After more than 50 years of continuous improvement, the controlled clinical trial has become the most effective way of determining the relative efficacy and toxicity of new drug therapies. This gold standard is, however, in jeopardy due to the expanding commend of the pharmaceutical industry. Mega-trials are not only paid for by the industry but also designed, carried-out, and analyzed by the industry. Because objectivity is at stake when industrial money mixes with the profession[9] it has been recently suggested to separate scientific research and the pharmaceutical industry. However, separation may not be necessary, and might be counterproductive to the progress of medicine. After all, pharmaceutical industry

has deserved substantial credits for developing important medicines, while other bodies including governments have not been able to develop medicines in the past 40 years, with the exception of one or two vaccines. Also, separation would mean that economic incentives are lost not only on the part of the industry but also on the part of the profession while both are currently doing well in the progress of medicine. Money *was* and *is* a major motive to stimulate scientific progress. Without economic incentives from industry there may soon be few clinical trials. Circumspection from independent observers during each stage of the trial has been recognized as an alternative for increasing objectivity of research and preventing bias.[3] In addition, tight control of study data, analysis, and interpretation by the commercial sponsor is undesirable. It not only raises the risk of biased interpretation, but also limits the opportunities for the scientific community to use the data for secondary analyses needed for future research.[4] If the pharmaceutical industry allows the profession to more actively participate in different stages of the trial, scientific research will be better served, and reasonable biological questions will be better answered. First on the agenda will have to be the criteria for adequate circumspection (table 2). Because the profession will be more convinced of its objective character, this allowance will not be counterproductive to the sales. Scientific research will be exciting again, confirming prior hypotheses, and giving new and sound ideas for further research.

Table 2. Criteria for adequate circumspection

1.	Disclosure of conflict of interests and the nature of it from each party involved
2.	Independent ethical and scientific assessment of the protocol
3.	Independent monitoring of the conduct of the trial
4.	Independent monitoring of data management
5.	Independent monitoring of statistical analysis including the cleaning-up of the data
6.	The requirement to publish even if data do not fit in the commercial interest of the sponsor.
7.	Requirement that interim analyses be performed by an independent group.

8. CONCLUSIONS

The controlled clinical trial, the gold standard for clinical research, is in jeopardy. The pharmaceutical industry rapidly expands its commend over clinical trials. Scientific rigor requires independence and objectivity. Safeguarding such criteria is hard with industrial sponsors, benefiting from favorable results, virtually completely in control. The objective of this chapter was to review flawed procedures jeopardizing the credibility of trials, and to look for possible solutions to the dilemma between sponsored industry and scientific independence.

Flawed procedures jeopardizing current clinical trials could be listed as follows. Industries, at least in Europe, are allowed to choose their own independent protocol review board prior to approval. The independent protocol review board approves protocols even if the research is beyond the scope of its expertise. Health facilities hosting multicenter trials are requested to refrain from scientific or ethic assessment of the trial. Trial monitors are often employees of industry. Data control is predominantly in the hands of the sponsor. Interim analyses are rarely performed by independent groups. The scientific committee of the trial consists largely of prominent but otherwise uninvolved physicians attached to the study. Analysis and report of the trial is generally provided by clinical associates at the pharmaceutical companies and, after a brief review, co-signed by prominent physicians attached to the study.

Possible solutions to the dilemma between sponsored industry and scientific independence could include the following. Circumspection from independent observers during each stage of the trial is desirable. In contrast, tight control of study data, analysis, and interpretation by the commercial sponsor is not desirable. If, instead, pharmaceutical industry allows the profession to more actively participate in different stages of the trial, scientific research will be better served, reasonable biological questions will be better answered, and, because the profession will be more convinced of the objective character of the research, it will not be counterproductive to the sales.

9. REFERENCES

1. Anonymous. International guidelines for good clinical practice. Ed: NEFARMA (Netherlands Association of Pharmaceutical Industries), Utrecht, Netherlands, 1997.
2. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948; 2: 769-82.
3. Cleophas TJ, Zwinderman AH, Cleophas TF. 2002 Statistics Applied to Clinical Trials, Self-assessment Book. Kluwer Academic Publishers, Boston, USA
4. Montaner JS, O'Shaughnessy MV, Schechter MT. Industry-sponsored clinical research: a double-edged sword. Lancet 2001; 358: 1893-5.
5. Diamant JC. The revised Declaration of Helsinki - is justice served. Int J Clin Pharmacol Ther 2002; 40: 76-83.

6. Sleight P, Yusuf S, Pogue J, Tsuyuki R, Diaz R, Probsfield J. Blood pressure reduction and cardiovascular risk in HOPE Study. Lancet 2001; 358: 2130-1.
7. 7.PROGRESS Collaborative Group. Randomised trial of a perindopril-based blood-pressure lowering regimen among 6105 individuals with previous stroke or transient ischaemic attack. Lancet 2001; 358: 1033-41.
8. Meyer FP, Cleophas TJ. Meta-analysis of beta-blockers in heart failure. Int J Clin Pharmacol Ther 2001; 39: 561-563 and 39: 383-8
9. Cleophas TJ, Zwinderman AH. Efficacy of HMG-CoA reductase inhibitors dependent on baseline cholesterol levels, secondary analysis of the Regression Growth Evaluation Statin Study (REGRESS). Br J Clin Pharmacol 2003; 56: 465-6.
10. Relman AJ, Cleophas TJ, Cleophas GI. The pharmaceutical industry and continuing medical education. JAMA 2001; 286: 302-4.

Appendix

T-Table: ν= degrees of freedom for t-variable, Q=area under the curve right from the corresponding t-value, 2Q tests both right and left end of the total area under the curve

ν	Q = 0.4 2Q = 0.8	0.25 0.5	0.1 0.2	0.05 0.1	0.025 0.05	0.01 0.02	0.005 0.01	0.001 0.002
1	0.325	1. 000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.547	5.841	10.213
4	.171	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.261	0. 700	1.372	1.812	2.228	2.764	3.169	4.144
11	.269	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.269	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.600	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1,316	1.708	2.060	2.485	2.787	3.450
26	.256	.654	1,315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1,314	1.701	2.052	2.473	2.771	3.421
28	.256	.683	1,313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.950	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

Chi-square distribution

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

F-distribution

df of denominator	2-tailed *P*-value	1-tailed *P*-value	Degrees of freedom (*df*) of the numerator												
			1	2	3	4	5	6	7	8	9	10	15	25	500
1	**0.05**	**0.025**	**647.8**	**799.5**	**864.2**	**899.6**	**921.8**	**937.1**	**948.2**	**956.6**	**963.3**	**968.6**	**984.9**	**998.1**	**1017.0**
1	0.10	0.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	249.3	254.1
2	**0.05**	**0.025**	**38.51**	**39.00**	**39.17**	**39.25**	**39.30**	**39.33**	**39.36**	**39.37**	**39.39**	**39.40**	**39.43**	**39.46**	**39.50**
2	0.10	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.46	19.49
3	**0.05**	**0.025**	**17.44**	**16.04**	**15.44**	**15.10**	**14.88**	**14.73**	**14.62**	**14.54**	**14.47**	**14.42**	**14.25**	**14.12**	**13.91**
3	0.10	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.63	8.53
4	**0.05**	**0.025**	**12.22**	**10.65**	**9.98**	**9.60**	**9.36**	**9.20**	**9.07**	**8.98**	**8.90**	**8.84**	**8.66**	**8.50**	**8.27**
4	0.10	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.77	5.64
5	**0.05**	**0.025**	**10.01**	**8.43**	**7.76**	**7.39**	**7.15**	**6.98**	**6.85**	**6.76**	**6.68**	**6.62**	**6.43**	**6.27**	**6.03**
5	0.10	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.52	4.37
6	**0.05**	**0.025**	**8.81**	**7.26**	**6.60**	**6.23**	**5.99**	**5.82**	**5.70**	**5.60**	**5.52**	**5.46**	**5.27**	**5.11**	**4.86**
6	0.10	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.83	3.68
7	**0.05**	**0.025**	**8.07**	**6.54**	**5.89**	**5.52**	**5.29**	**5.12**	**4.99**	**4.90**	**4.82**	**4.76**	**4.57**	**4.40**	**4.16**
7	0.10	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.40	3.24
8	**0.05**	**0.025**	**7.57**	**6.06**	**5.42**	**5.05**	**4.82**	**4.65**	**4.53**	**4.43**	**4.36**	**4.30**	**4.10**	**3.94**	**3.68**
8	0.10	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	2.94
9	**0.05**	**0.025**	**7.21**	**5.71**	**5.08**	**4.72**	**4.48**	**4.32**	**4.20**	**4.10**	**4.03**	**3.96**	**3.77**	**3.60**	**3.35**
9	0.10	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.89	2.72
10	**0.05**	**0.025**	**6.94**	**5.46**	**4.83**	**4.47**	**4.24**	**4.07**	**3.95**	**3.85**	**3.78**	**3.72**	**3.52**	**3.35**	**3.09**
10	0.10	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.73	2.55
15	**0.05**	**0.025**	**6.20**	**4.77**	**4.15**	**3.80**	**3.58**	**3.41**	**3.29**	**3.20**	**3.12**	**3.06**	**2.86**	**2.69**	**2.41**
15	0.10	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.28	2.08
20	**0.05**	**0.025**	**5.87**	**4.46**	**3.86**	**3.51**	**3.29**	**3.13**	**3.01**	**2.91**	**2.84**	**2.77**	**2.57**	**2.40**	**2.10**
20	0.10	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.07	1.86
30	**0.05**	**0.025**	**5.57**	**4.18**	**3.59**	**3.25**	**3.03**	**2.87**	**2.75**	**2.65**	**2.57**	**2.51**	**2.31**	**2.12**	**1.81**
30	0.10	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.88	1.64
50	**0.05**	**0.025**	**5.34**	**3.97**	**3.39**	**3.05**	**2.83**	**2.67**	**2.55**	**2.46**	**2.38**	**2.32**	**2.11**	**1.92**	**1.57**
50	0.10	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.73	1.46
100	**0.05**	**0.025**	**5.18**	**3.83**	**3.25**	**2.92**	**2.70**	**2.54**	**2.42**	**2.32**	**2.24**	**2.18**	**1.97**	**1.77**	**1.38**
100	0.10	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.62	1.31
1000	**0.05**	**0.025**	**5.04**	**3.70**	**3.13**	**2.80**	**2.58**	**2.42**	**2.30**	**2.20**	**2.13**	**2.06**	**1.85**	**1.64**	**1.16**
1000	0.10	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.52	1.13

Paired non-parametric test: Wilcoxon signed rank test, the table uses smaller of the two ranknumbers

N pairs	P<0.05	P<0.01
7	2	0
8	2	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	19

Unpaired non-parametric test: Mann-Whitney test. Table uses difference of added up rank numbers between group 1 and group 2

P<0.05 levels

n_2 ↓ / n_1 →	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5				15										
6			10	16	23									
7			10	17	24	32								
8			11	17	25	34	43							
9		6	11	18	26	35	45	56						
10		6	12	19	27	37	47	58	71					
11		6	12	20	28	38	49	61	74	87				
12		7	13	21	30	40	51	63	76	90	106			
13		7	14	22	31	41	53	65	79	93	109	125		
14		7	14	22	32	43	54	67	81	96	112	129	147	
15		8	15	23	33	44	56	70	84	99	115	133	151	171
16		8	15	24	34	46	58	72	86	102	119	137	155	
17		8	16	25	36	47	60	74	89	105	122	140		
18		8	16	26	37	49	62	76	92	108	125			
19	3	9	17	27	38	50	64	78	94	111				
20	3	9	18	28	39	52	66	81	97					
21	3	9	18	29	40	53	68	83						
22	3	10	19	29	42	55	70							
23	3	10	19	30	43	57								
24	3	10	20	31	44									
25	3	11	20	32										
26	3	11	21											
27	4	11												
28	4													

Unpaired non-parametric test: Mann-Whitney test. Table uses difference of added up rank numbers between group 1 and group 2

P<0.05 levels

n_2 ↓ / n_1 →	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5				15										
6			10	16	23									
7			10	17	24	32								
8			11	17	25	34	43							
9		6	11	18	26	35	45	56						
10		6	12	19	27	37	47	58	71					
11		6	12	20	28	38	49	61	74	87				
12		7	13	21	30	40	51	63	76	90	106			
13		7	14	22	31	41	53	65	79	93	109	125		
14		7	14	22	32	43	54	67	81	96	112	129	147	
15		8	15	23	33	44	56	70	84	99	115	133	151	171
16		8	15	24	34	46	58	72	86	102	119	137	155	
17		8	16	25	36	47	60	74	89	105	122	140		
18		8	16	26	37	49	62	76	92	108	125			
19	3	9	17	27	38	50	64	78	94	111				
20	3	9	18	28	39	52	66	81	97					
21	3	9	18	29	40	53	68	83						
22	3	10	19	29	42	55	70							
23	3	10	19	30	43	57								
24	3	10	20	31	44									
25	3	11	20	32										
26	3	11	21											
27	4	11												
28	4													

INDEX