Ton J. Cleophas
Aeilko H. Zwinderman

Analysis of Safety Data of Drug Trials

An Update

EXTRAS ONLINE

Springer

Analysis of Safety Data of Drug Trials

Ton J. Cleophas • Aeilko H. Zwinderman

Analysis of Safety Data of Drug Trials

An Update

Springer

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

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

Additional material to this book can be downloaded from http://extras.springer.com.

ISBN 978-3-030-05803-6 ISBN 978-3-030-05804-3 (eBook)
https://doi.org/10.1007/978-3-030-05804-3

Library of Congress Control Number: 2018966807

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

Preface

In 2010, the fifth edition of the textbook *Statistics Applied to Clinical Studies*, Springer, Heidelberg, Germany, was published by the authors, and over a million copies have been sold. The primary objective 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 for the purpose is to present summaries of prevalences of adverse effects and their 95% confidence intervals. In order to estimate the probability that the differences between treatment and control group did not occur merely by chance, a statistical test can be performed. In the past few years, this pretty crude method has been supplemented and, sometimes, replaced with more sophisticated and better sensitive methodologies, based on machine learning clusters and networks, and multivariate analyses. And so, it is time that an updated version of safety data analysis was published.

For the statistical analysis of safety data, better-fit methods are, thus, available, and this is fine. There is, however, another important topic brought forward in connection with safety data analyses but, maybe, also relevant to the statistical analysis of clinical trials in general. It includes novel insights into hypothesis testing, favoring the alternative hypothesis over the null hypothesis.

Also, the issue of *dependency* needs to be addressed. Adverse effects may be either dependent or independent of the main outcome. For example, an adverse effect of alpha blockers is dizziness, and this occurs independently of the main outcome "alleviation of Raynaud's phenomenon." In contrast, the adverse effect "increased calorie intake" occurs with "increased exercise," and this adverse effect is very dependent on the main outcome "weight loss." Random heterogeneities, outliers, confounders, and interaction factors are common in clinical trials, and all of them can be considered as kinds of adverse effects of the dependent type. Random regressions and analyses of variance, high dimensional clusterings, partial correlations, structural equations models, and other Bayesian methods are helpful for their analysis.

The current edition was written for non-mathematicians, particularly medical and health professionals and students. It provides examples of modern analytic methods so far largely unused in safety analysis. All of the 16 chapters have two core characteristics, First, they are intended for current usage, and they are particularly concerned with that usage. Second, they try and tell what readers need to know in order to understand and apply the methods. For that purpose, step-by-step analyses of both hypothesized and real data examples will be given. Each chapter can be studied as a stand-alone.

Sliedrecht, The Netherlands — Ton J. Cleophas
Amsterdam, The Netherlands — Aeilko H. Zwinderman

Contents

Part I The Analysis of Independent Adverse Effects

Chapter 1
General Introduction

Abstract The current chapter reviews the history of adverse effects of modern medicines from the era of pharmageddon to the current era of precision medicine. Adverse drug effects are classified significant, if their 95% confidence interval is significantly different from zero or control. They are classified dependent, if they are significantly dependent not only on the efficacy variable of the study but also on the main outcome variable. They may be harder to recognize and may go undetected if not special methods for assessment have been used.

A brief review is given of the detection and assessment of different types of dependent adverse effects.

Mechanisms of dependency may be:

causally,
pharmacologically,
through interaction,
through a subgroup mechanism,
through a pleiotropic drug effect,
through carryover effect,
through a categorical effect,
through confounding.

Particular attention will be given to structural equation modeling for the purpose of a rapid identification of types of relationships.

Keywords Pharmageddon · Precision medicine · Significant and insignificant adverse effect · Dependent and independent adverse effect · Structural equation modeling

1 Introduction, Pharmageddon, Efficacious Treatments

In 1974 an alarming article entitled Medical Nemesis was written by the New York general practitioner Ivan Illitch (Lancet 1974; i: 918–21). It described how, at that time, modern medicines had dramatic sickening power, and had become a major

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_1

threat to health rather than the opposite. The term Nemesis, was used by Illitch, because, in ancient Greece, it was the goddess who severely punished arrogance before the gods. The article contributed to the development of the concept "pharmageddon", an amalgamation of pharmacy and armageddon, where armageddon is a term used to describe the gathering of armies in the end of times. Medicines were thought to be like arms, equally destructive. All of this was a consequence of years of novel medicines that were very unsafe as documented soon after approval. In 2002 the Br Med J published a reprint of the article in memory of the author. But, then, pharmageddon was almost past and novel efficacious medicines were being developed. In 20 subsequent years the 30% survival from cancer changed and improved to over 70%. Also better drugs for cardiovascular diseases were being invented like powerful anticholesterol and anticoagulant agents. Obviously, we had started to have medicines that worked.

Did this improvement of efficacy also alter the importance and approach to safety analyses of new drugs? According to Bayer (Columbia University New York, N Engl J Med 2015; 373: 499–502) modern clinical medicine has contributed enormously to our ability to treat and cure sick people. However, at the population level benefits of the least advantaged are missing. In the US out of all countries, life expectancies even sunk, and they continue to do so today. Currently, two types of medical treatments can be distinguished. First, precision medicine, focusing on detecting and curing disease at the individual level. Second, health care at the population level. President Obama's State of the Union, 2015, exulted at the first. Varmus, director of the National Cancer Institute, and Collins, director of the National Institute of Health, addressed and expressed their worries about the second, and called for a broad research program for building an evidence-base for guiding a better clinical practice in the future (N Engl J Med 2015; 372: 793–795). This research program should be, particularly, concerned with prevention, and elimination of risk factors, which are the adverse effects of modern treatments. And so, despite the good news about efficacious treatments, safety analysis and focus on risk factors of treatments and health in general are still relevant today, maybe even more so than before given the continual increase of drug consumption.

2 Some Terminology

Bayesian Networks
A structural equation model is currently commonly named a Bayesian network, otherwise called a DAG (directed acyclic graph). It is a probabilistic graphical model of nodes (the variables) and connecting arrows presenting the conditional dependencies of the nodes.

Categorical Predictors
Still another significant adverse effect may occur, if you replace an insignificant predictor with a categorical one.

Causal Adverse Effects, Path Analysis, Partial Correlation
Sometimes a subgroup effect is most probably a causal effect. Path analysis can tell causal and non-causal effects apart. If in a partial correlation analysis of a three step path analysis the second path is held constant, the correlation between factor 1 and 3 may or may not disappear. If not, it must have been causal.

Defining Adverse Effect
In this edition we will use the term adverse effect as basic term covering all kinds of unexpected and expected effects in clinical trials if they are not the protocol's main outcome.

EUDIPHARM (European College of Pharmaceutical Medicine)
Academic College sponsored by the European Community Socrates Project with headquarters in Lyon France providing a doctorate in pharmaceutical medicine and mainly involved in all aspects of pharmacovigilance.

FDA's Final Rule on Expedited Safety Reporting
In 2011 a final rule for expedited reporting of serious adverse events took effect in the USA for studies conducted under an Investigational New Drug application (see Wittes et al. Stat Biopharmaceutic Res 2015; 7:3: 174–190). Specific statistics included: one sided 80% confidence interval of adverse event rates observed versus control should not include 0, relative risk compared to control should be >2.

Guidelines for Good Clinical Practice
Guidelines writtten by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, in July 2014.

Higher Order Partial Correlations and Higher Order Loglinear Models
The more complex your statistical analyses, the more unpredicted effects will be encountered. Notorious examples are the higher order partial correlation models and the higher order loglinear models for which modern statistical software usually provides ample modules. You may call it explorative research of low scientific validity, but we live in an era of big, and, therefore, powerful data. And explorative research is currently the main stream in data mining and machine learning analyses, and results are currently increasingly taken serious unlike in the past, and rightly so.

Pharmacovigilance
Pharmacovigilance literally means drug safety. The term is used to indicate the pharmacological science of detection, assessment, monitoring, and prevention of adverse effects from pharmaceutical products.

Pleiotropy
Sometimes an interaction effect in the data are most probably due to an effect of pleiotropy. With pleiotropy a single gene is responsible for multiple patient characteristics. Bayesian networks between variables often give rise to this form of unexpected adverse effect either detrimental or beneficial.

Safety Signal
Data information, suggesting a new causal association between a medicine and an adverse effect.

Safety Signal Detection
A field defined as the methodology for summarizing information about safety levels of novel food and drug compounds from multiple studies.

Side Effects and Adverse Effects
Also terminologies are not uniformly applied. Side effects and adverse effects are synonymous, but, in practice, the former is mainly used for the less severe and the latter for the more severe effects. A side effect is something for which reassurance will be adequate, but adverse effects require serious assessment. But things are more complex than that.

Side Effect Rating Scales
Side effects in drug trials, although of unequivocal importance, are usually assessed in a pretty unstructured way. Recently, some side effect rating scales have been proposed, for example, the GASE (Generic assessement of side effects in clinical trials), UKU (Udvalg for kliniske undersogelser side effect rating scale from the Norwegian Directorate of Health), FDA (Food Drug Administration) regulations of common drug side effects, the SAFTEE (Systematic assessment for treatment emergent events from the American National Institute of Mental Health) rating scale and more. However, consensus of how to analyze listings from such ratings in a statistically meaningful way, otherwise called the scientific method, is lacking.

Structural Equation Models (SEMs)
In clinical efficacy studies the outcome is often influenced by multiple causal factors, like drug – noncompliance, frequency of counseling, and many more factors. Structural equation modeling (SEM) was only recently formally defined by Pearl 2000. This statistical methodology includes

1. factor analysis,
2. path analysis,
3. regression analysis.

An SEM model looks like a complex regression model, but it is more. It extends the prior hypothesis of correlation to that of causality, and this is accomplished by a network of variables tested versus one another with standardized rather than non-standardized regression coefficients.

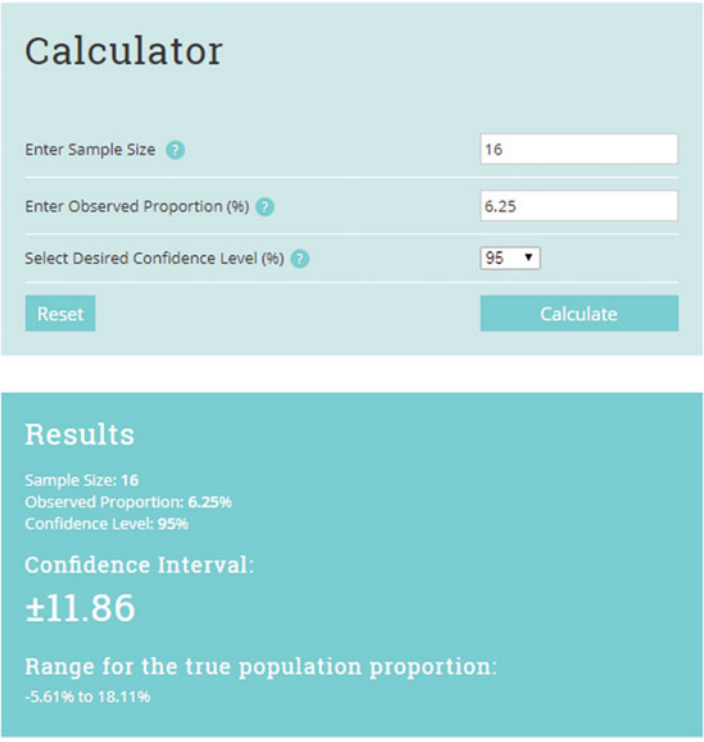
Subgroup Effects
Side effects serious or not may be subgroup effects. They may be confounders if they are present in treatment and control groups, or interaction factors if present only in one of the two groups. They may also be caused be outlier clusters in the data. Side effects may be random effects, which are unexpected subgroup effects. If a random effect statistical analysis is positive, then the random effect will be partly responsible for the overall effect in a study.

3 Significant and Insignificant Adverse Effects in Clinical Trials

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 clinical trial. An effective approach for the purpose is to compute summaries of prevalences of adverse effects and their 95% confidence intervals.

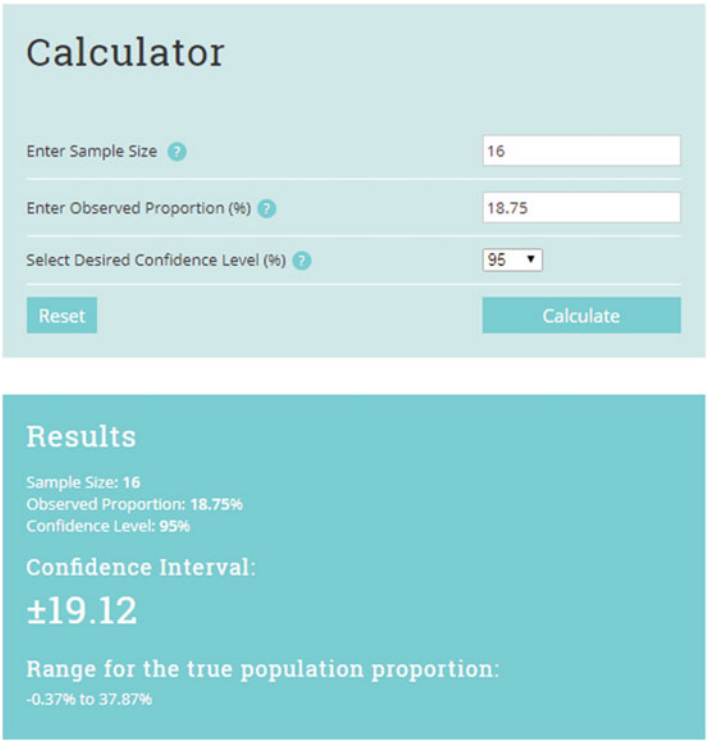
Significantly Different from Zero

In order to estimate whether the 95% confidence interval of an adverse effect is significantly different from a prevalence of zero, we will use the confidence interval calculator for proportions from Allto Ltd. T/asa Allto Consulting Leeds UK (info@allto.co.uk).

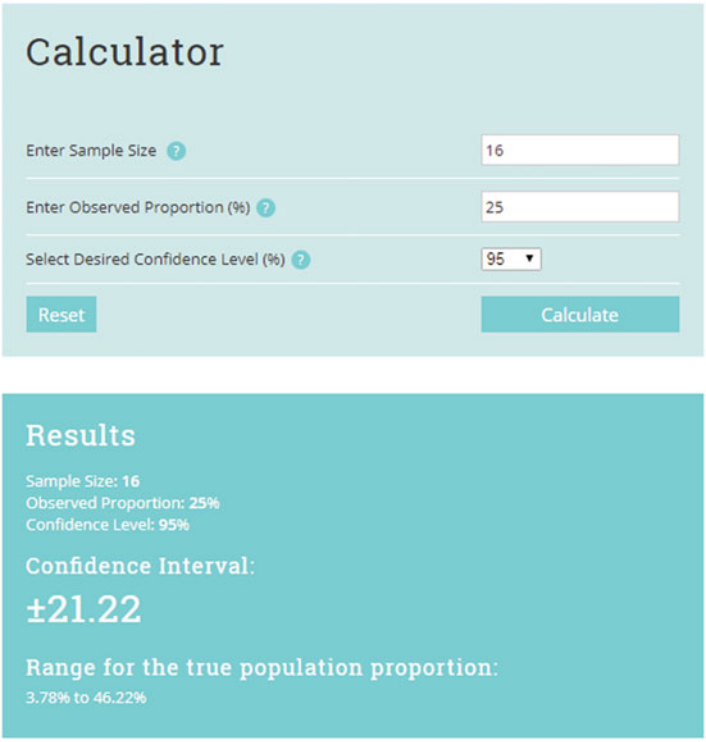


The above graph shows that, if in a sample of 16 patients only one patient suffers from a particular adverse effect, then the difference from a prevalence of zero will

not be statistically significant. The underneath graph shows that, if three patients suffer from an adverse effect, the 95% confidence interval of this proportion will be between −0.37% and 37.87%. The left end of the 95% confidence interval will be close to zero but does not cross the zero prevalence cut-off.



The underneath graph shows the results with a proportion of patients with the adverse effect being 25%. Now the left end of the 95% confidence interval is larger than a prevalence of zero. The interval is between 3.78% and 46.22%, and does not include 0% anymore. This means, we have <5% chance that zero is included and, so, our result is significantly different from zero. However, a p-value of 5% is not powerful and the chance of type I errors of finding no difference where there is one is at least 5%. Nonetheless, it is usually stated that with 4 out of 16 patients having a particular adverse effect (= 25%) would mean, that a significant adverse effect is in this sample.



Significantly Different from Control

Often in clinical trials not a single sample effect is tested against zero, but rather a treatment and control group are compared, and, in order to estimate whether one treatment has more adverse effects than the other, we will try and test, whether the difference of proportions in either of the treatment groups are not merely chance but statistically significant. For that purpose various statistical tests are available, for example:

Z-Tests
Chi-Square tests
Fisher exact tests
Multiple Chi-Square tests
McNemar tests
Multiple McNemar tests and Cochrane Q-Tests
Odds Ratio tests
McNemar Odds Ratio tests

Log Likelihood Ratio tests
Cox regression tests
Logistic regression tests
Hazard Ratio tests
Bayesian t-, anova-, chi-square tests

Step by step analyses using the above tests will be covered in the Chap. 2. In the past few years, this pretty crude method has been supplemented, and, sometimes, replaced with more sophisticated and better sensitive methodologies, based on machine learning clusters and networks, and multivariate analyses. And so, it is time that an updated version of safety data analysis was published. Updated safety data analyses are the main subject of this edition, and they will be reviewed in the remainder of the Chaps.

4 Independent and Dependent Adverse Effects

Terminologies of adverse effects are inconsistent and pretty confusing. For example, the term adverse effects is usually used for adverse drug events (Guidelines for Good Clinical Practice, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, July 2014). and refers to injury at the time a drug is used, and they may be causal or not. If causal they will be named adverse drug reactions. This is different from side effects, because side effects may also be beneficial. The field of pharmacovigilance is involved in the study of adverse drug reactions. Adverse drug events are assumed to be causal and, if dose dependent, we will call it type A, if not, we will call it idiosyncratic. They are furthermore classified many ways, for example, according to severities, locations, mechanisms, paradoxical reactions, levels of polypharmacy, iatrogenesis, synergisms (interactions), etc. In order to try and reduce the inconsistency of terminologies, our institution at the Claude Bernard University in Lyon, called the European College of Pharmaceutical Medicine EUDIPHARM, has decided, already 20 years ago, to slightly adapt and minimize terminologies, and choose the terms independent and dependent adverse effects, covering all of the possible adverse drug reactions. It means, that adverse effects may be either dependent or independent of the main outcome. For example, an adverse effect of alpha blockers is dizziness, and this occurs independently of the main outcome “alleviation of Raynaud ‘s phenomenon”. In contrast, the adverse effect “increased calorie intake” occurs with “increased exercise”, and this adverse effect is very dependent on the main outcome “weight loss”. All of the methodologies reviewed in the Chaps. 2, 3, 4, 5, 6, 7, 8 and 9 are for analyzing independent adverse effects.

5 A Brief Review of Methods for Detection and Assessment of *Independent* Adverse Effects

The methodologies reviewed in the Chaps. 2, 3, 4, 5, 6, 7, 8 and 9 are for analyzing independent adverse effects. We will start with explaining traditional and more modern statistical tests for assessing the presence of statistically significant and insignificant adverse effects, and we will use step by step analyses of data examples (Chap. 2). Incidence ratios, reporting ratios and safety signals based on multiple criteria may be used instead of adverse effects (Chap. 3). Different classes of severity may require different statistical analyses (Chap. 4).

Forest plots (Chap. 5) were originally invented to visualize in meta-analyses the main effects of the separate studies included, but they are also helpful in clinical studies to quantitatively and qualitatively analyze the presence of common adverse effects. In the Chap. 6 a systematic assessment of qualitative adverse effects as commonly observed are given, and graphs of the odds of patients with adverse effects having had a medication or not, and ratios of those odds in the form of forest plots are used for clarification.

Computer files of clinical data are often complex and multi-dimensional, and they may be hard to statistically test. Instead, visualization processes including both binary and continuous data may be helpful. Graphics using different visualization methods as available in current data mining programs will be used for the purpose. In the Chap. 6 we will apply for example the Konstanz Information Miner (KNIME) and WEKA (Waikato University New Zealand) miner, widely approved and appreciated free machine learning software packages on the Internet since 2006.

More longitudinal studies often include repeated outcome measures, and such studies greatly benefit from adjustments for time effects. Mixed linear models will be particularly adequate for the purpose, and provides better power than traditional repeated measures analysis of variance, because within subject differences receive fewer degrees of freedom. In the Chap. 7 it will be shown that in this way a better sensitivity is left in the analysis to demonstrate differences between subjects.

Benefit risk assessments are more relevant with respect to the safety of new drugs than anything else, because one may cancel out the other, and benefit risk assessment is according to the FDA (Chap. 8) the single basis for regulatory review of new drugs. Unfortunately, benefit risk ratios are currently assessed in a colloquial rather than analytical way. This edition will, however, demonstrate that an analytical assessment including computed confidence intervals of such ratios is not impossible, and that this assessment will be a major aim of clinical research in the near future.

The presence of adverse effects in current equivalence, inferiority, and superiority trials have to be assessed differently from the traditional approach of null hypothesis testing. This is because the presence of adverse effects is here not confirmed by the rejection of some null hypothesis but rather it will be confirmed if a priori defined boundaries are met. The Chap. 9 will give various examples.

6 A Brief Review of Methods for Detection and Assessment of *Dependent* Adverse Effects

Adverse effects may be either dependent or independent of the main outcome. How do we assess dependent adverse effects. Drug induced *independent* adverse effects is the main subject of the Chaps. 2, 3, 4, 5, 6, 7, 8 and 9, and they are commonly and easily observed in clinical trials. However, drug-induced *dependent* adverse effects are also pretty common. But they may be harder to recognize, and may go undetected, if not special methods of assessment are applied. The results of the trial may be meaningless without proper detection and adjustment. For example, a significant interaction effect between genders on the outcome may obscure treatment efficacies, if one gender receives more often the treatment it better responds to than the other. Separate analyses for separate genders can adjust and correct this deleterious adverse effect.

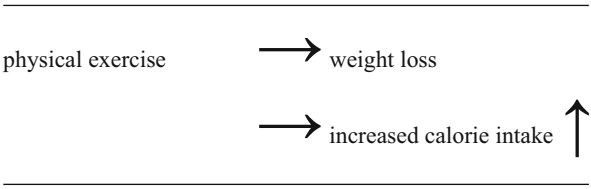
Also random heterogeneities, outlier data, confounders, interaction factors are not uncommon in clinical trials, and all of them can, equally so, be considered as kinds of adverse effects of the dependent type. Random regressions and analyses of variance, high dimensional clustering, partial correlations, structural equations models, and other Bayesian methods are helpful for their analysis. We should add, that, unlike independent adverse effects, dependent adverse effects, as they are not always easy to identify, require advanced methodologies for their detection. These methodologies, usually, make use of different forms of regression analyses, and general principles of regression analyses are, therefore relevant to keep in mind. Regression analysis uses predefined mathematical models, and, then, applies the data to compute the best fit parameters, like the best fit lines, exponential curves, curvilinear curves (those that have the shortest distance from the data), and, subsequently, it tests, how far distant from the curve the data are. A significant correlation between the y- and x- data means that the y-data are closer to the model than will happen with random sampling (i.e., by chance). The distances are, finally, statistically tested with pretty simple statistical tests like t-tests or analyses of variance. The "model principle" is wonderful for fitting data, but, it is, at the same time, its largest limitation, because it's often no use forcing nature into a mathematical model. This edition will address and explain many methods for detecting and analyzing dependent adverse effects. Virtually all of these methods include elements of regression analyses.

For a better understanding of mechanisms responsible for dependent adverse effects we need to demonstrate some kind of relationships between the dependent adverse effect factor and the study outcome. A few examples will be given underneath, but more detailed computational analyses will be given in the Chaps. 10, 11, 12, 13, 14, 15 and 16.

7 Examples of Causal Relationships Between Dependent Adverse Effect and Outcome

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With causal relationships structural equation models and Bayesian networks are adequate for the purpose.

Example 1

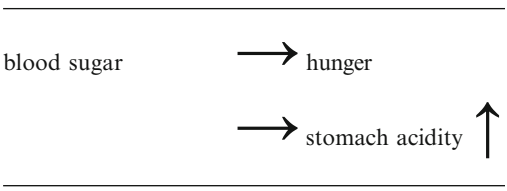


The main outcome of physical exercise in a study is weight loss. However, increased calorie intake is also caused by physical exercise, and this counteracts the effect on weight loss. And so, increased calorie intake is an adverse effect of physical exercise on the outcome weight loss. This adverse effect is significantly related to the outcome weight loss, and, so, we will call it a dependent causal adverse effect of physical exercise on the outcome weight loss. The above graph is a kind of structural equation model where three arrows indicating a positive statistically significant correlation between.

(1) physical exercise and weight loss
(2) physical exercise and increased calorie intake
(3) increased calorie intake and weight loss.

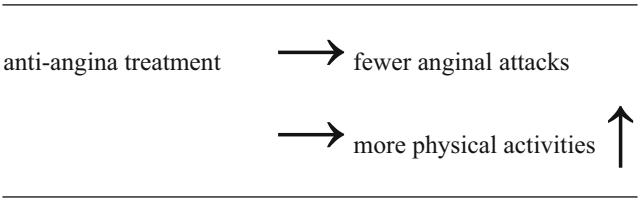
Often, a structural equation model is named a Bayesian network, because the Bayes equation "prior odds x Bayes factor = posterior odds" is textually and conceptionally very similar to a structural equation model with arrows and path statistics.

Example 2



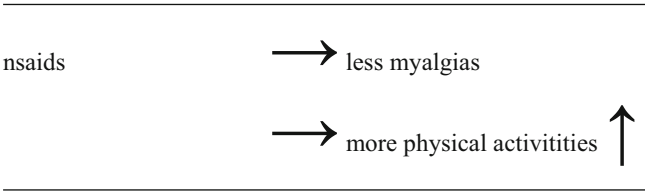
Stomach acidity is a dependent adverse effect of blood sugar on hunger, because it changes the outcome of the study, hunger, the pattern 'blood sugar → stomach acidity → hunger" is called a structural equation model where the arrows used indicate standardized regression coefficients rather than non-standardized ones. Structural equation models are the basis of multistep path analysis, partial correlation models and Bayesian networks. More details will be covered in the Chap. 10.

Example 3



The presence of more physical activities is a dependent adverse effect of anti-angina treatments on the outcome "fewer anginal attacks", because it changes the magnitude of the outcome.

Example 4



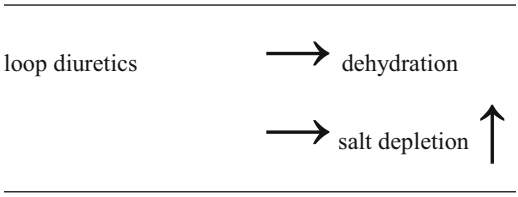
The presence of more physical activities is a dependent adverse effect of nsaids treatment (non-steroidal anti-inflammatory drugs) on the outcome "less myalgias", because it changes the magnitude of the outcome. The presence of more physical activities is a dependent adverse effect, because it changes the output "less myalgias" (nsaids are nonsteroidal anti-inflammatory drugs).

8 Examples of Pharmacological Mechanisms Between Dependent Adverse Effect and Outcome

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an

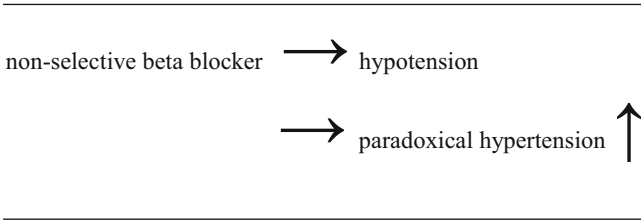
intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With pharmacological mechanisms traditional t-test for continuous and chi-square tests for binary data are adequate for the purpose.

Example 1



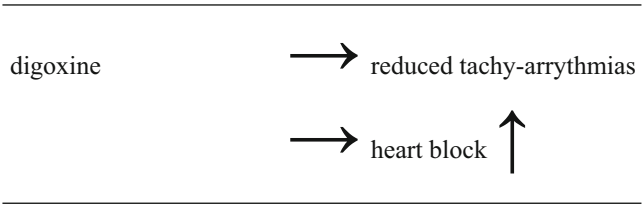
The presence of salt depletion is a dependent adverse effect, because it changes the outcome dehydration.

Example 2



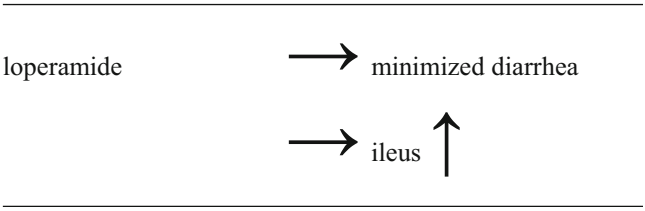
The presence of paradoxical hypertension is a dependent adverse effect, because it changes the outcome hypotension.

Example 3

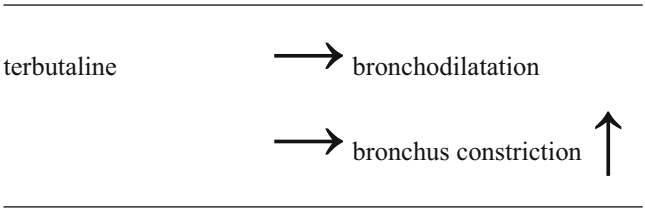


The presence of heart block is a dependent adverse effect, because it changes the outcome reduced tachy-arrythmias.

Example 4



Example 5
The presence of ileus is a dependent adverse effect, because it changes the "outcome minimized diarrhea".

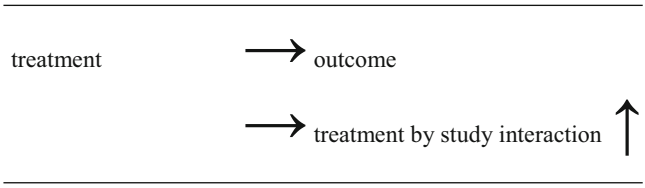


The presence of bronchus constriction is a dependent adverse effect, because it changes the outcome "bronchodilatation".

9 Example of Interaction Between Dependent Adverse Effect and Outcome

Causal relationships, pharmacological mechanisms, interactions (Chap. 13), subgroup mechanisms (Chap. 14), carryover effects from previous treatments (Chap. 12), pleiotropic drug mechanisms, categorical factors (Chap. 11), confoundings (Chap. 13), may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With interactions t-tests, analyses of variance, regressions and random effects tests are adequate for the purpose.

Example

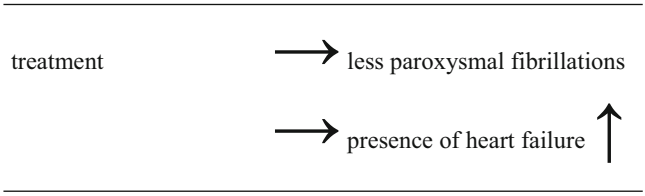


The presence of treatment by study interaction is a dependent adverse effect, because it changes the outcome, this is a random rather than fixed effect, and random effect analysis is required (Chap. 11).

10 Example of Subgroup Mechanism Between Dependent Adverse Effect and Outcome

A dependent adverse effect must be significantly related to the outcome. With subgroup mechanisms regression analyses are adequate for the purpose (Chap. 14).

Example

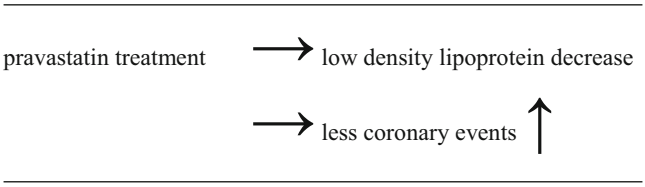


The presence of heart failure is a dependent adverse effect, because it increases the numbers of paroxysmal atrial fibrillations.

11 Examples of Pleiotropic Drug Mechanism Between Dependent Adverse Effect and Outcome

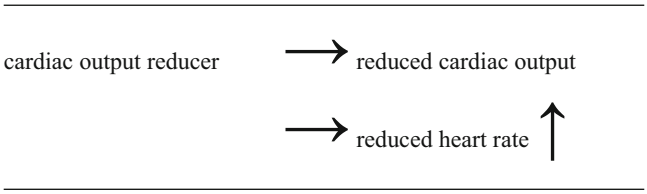
Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With pleiotropic drug mechanisms Bayesian networks are adequate for significance testing (Chap. 10).

Example 1



Less coronary events is a pleiotropic effect of randomized treatment and thus a dependent adverse effect.

Example 2

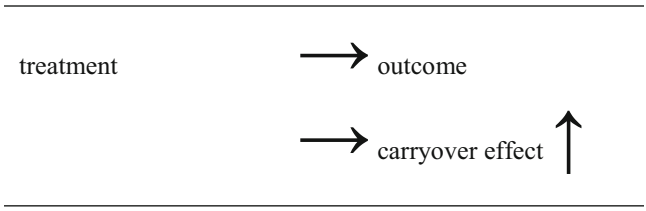


Reduced heart rate is a dependent adverse effect, because it changes the output cardiac output.

12 Example of a Carryover Mechanism Between Dependent Adverse Effect and Outcome

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With carryover effects traditional t-test, chi-square tests, Fisher exact tests are adequate for significance testing.

Example



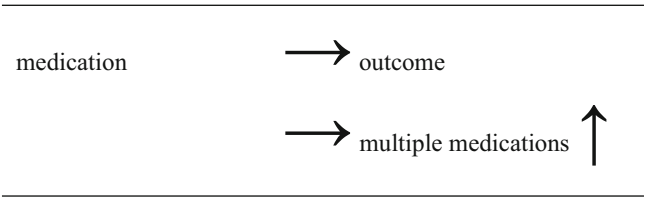
As an example of a dependent adverse effect carryover effect in crossover studies will be used. Carryover effect, otherwise called treatment by period interaction, is a major adverse effect of crossover studies. If the effect of a treatment carries on into the second period of treatment, then the measured response to the second period of treatment is changed. This carryover effect can be considered a dependent adverse effect (Chap. 12). Carryover effect is a dependent adverse effect, because it changes the outcome of the study. The above arrows often indicate positive Pearson

correlation coefficients, but other statistics are possible. In many cases, particularly with Bayesian networks, it also means an assumed causal relationship. Underneath more examples of dependent adverse effects are given.

13 Example of a Categorical Rather than Ordinal Mechanism Between Dependent Adverse Effect and Outcome

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With categorical mechanisms categorical regressions are adequate for significance testing.

Example

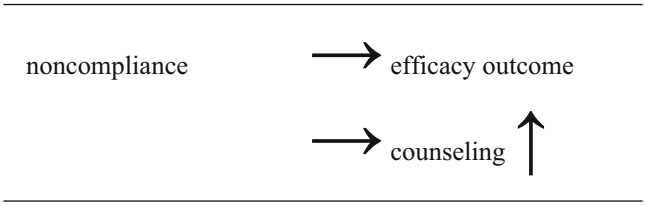


"Multiple medications"is a dependent adverse effect, because it changes the outcome of the study, for the purpose a continuous variable has to be replaced with a categorical one (Chap. 11).

14 Example of Confounding Between Dependent Adverse Effect and Outcome

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to intervention but also the outcome. With confoundings, subclassification, regression analysis, propensity scores are adequate for significance testing.

Example



Counseling is a dependent adverse effect, because it changes the efficacy outcome of the study, if it only changes the outcome of one treatment modality, we will call the adverse effect interaction, if that of two treatment modalities, we will call it confounding (Chap. 13).

15 Discussion

This chapter started with the term pharmageddon, an amalgamation of pharmacy and armageddon. In the early seventies medicines were thought to be like arms, equally destructive. This was a consequence of years of novel medicines that were very unsafe as documented soon after approval. Soon after we started to have efficacious medicines, but although particular efficacious in precision medicine, an expensive approach, they were not so for health care at the population level. Even today life expectancies are sinking, and, so, despite the good news about efficacious treatments, safety analysis and focus on risk factors of treatments and health in general are relevant today. Safety analysis in clinical research mainly involves the search for and study of adverse effects of treatments. Adverse effects may be statistically significantly present in a treatment group versus zero, but the term significant adverse effect is mainly applied if it is significantly more present in the treatment group than it is in the control group. Multiple testing is obviously an issue here, and adjustments accounting type I errors are in place. Adverse effects may be independent or dependent of the outcome. Brief reviews of the methods for detection and assessment of both independent and dependent adverse effects are given. Particularly dependent adverse effects are a tricky class that may easily go undetected, and require special expertise as well as special methods of analyses that will be the main subject of this edition. Mechanisms responsible for dependent adverse effects include pharmacological mechanisms, interactions, subgroup mechanisms, pleiotropic drug mechanisms, carryover mechanisms from prior treatments, categorical rather than ordinal mechanisms, confounding, outlier clusters, hierarchical and higher order effects.

The Chaps. 2, 3, 4, 5, 6, 7, 8 and 9 will review the analysis of independent adverse effects, while the Chaps. 10, 11, 12, 13, 14, 15 and 16 will particularly address the analysis of dependent adverse effects.

16 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Part I
The Analysis of Independent Adverse Effects

Chapter 2
Statistically Significant and Insignificant Adverse Effects

Abstract The current chapter reviews many statistical hypothesis tests, adequate for assessing prevalence (proportion of a population) and rate data (frequency of events per time unit) versus control or versus zero.

Only independent adverse effects are assessed here, and they are tested for statistically significant presence.

Both attention is given to paired and unpaired data, and particular attention is given to explicit time dependent Poisson methods, as well as log likelihood ratio tests, that provide better power than traditional tests for the purpose.

Bayesian crosstabs may be prone to overdispersion, but in the example given no adjustment was needed.

Methods for analyzing contingency tables larger than 2×2 are given, and longitudinal data like survival data and Cox regressions are used for addressing times to event and computing hazard ratios of adverse effects.

Keywords Z-tests · Chi-square-tests · Pocket Calculator Method · Fisher Method · Paired Unpaired Proportions · Cochran Test · Survival Analysis · Odds Ratios · Log Likelihood Ratios · Logistic Models · Poisson models · Cox Models · Bayesian Crosstabs

1 Introduction

An effective approach to the analysis of adverse effects is to present summaries of prevalences. The prevalence is synonymous to the proportion of a sample with an adverse effect. Prevalences are commonly tabled with confidence intervals (CIs), e.g., 95% CIs, that, under the assumption of a normal distribution, are estimated as follows

$$\pm 1.96 \ \sqrt{p(1-p)/n},$$

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_2

where p = the proportion of patients with an adverse effect and n is the magnitude of the sample. An example of common adverse effects are given underneath. In the above equation 1.96 is often rounded off to 2.0. Calculators for confidence intervals are widely available at the internet.

	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

The numbers in the table relate to the numbers of patients showing a particular adverse effect. Some questions were not answered by all patients. Particularly, sleepiness occurred differently in the two groups: 33% in the 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 for Testing Significance of Difference of Two Unpaired Proportions

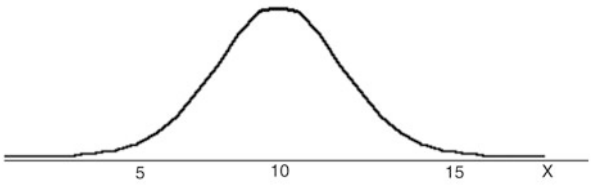
Many methods exist to analyze two unpaired proportions, like odds ratios analysis and logistic regression, but here we will start by presenting the four most common methods for that purpose. Using the sleepiness data from above we construct a 2×2 contingency table:

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

2.1 Method 1, Z – Test

We can test the significance of difference similarly to the method used for testing continuous data. 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 equation $SD = \sqrt{(p(1-p))}$. Unlike the SD for continuous data given by the equation $SD = \sqrt{\left(\frac{\sum (x-\bar{x})^2}{n-1}\right)}$, 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 the underneath example.

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). The figure below, with a frequency distribution of numbers of sleepy people observed in multiple samples of 15 patients from the same population, 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. And, so, according to the curve below SD = p (1−p) should be close to the truth.



Note that, for null-hypothesis-testing, standard error (SE or SEM) 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, is equal to √ p(1−p)/n . It

is relevant to mention here the "plus four rule". Instead of $\sqrt{}$ p(1−p)/n an adjusted standard error is often used that better fits lopsided data and small data and it goes like this:

p is replaced with (counts of successes +2 / counts of all observations + 4) = p′,
$\sqrt{}$ p(1−p)/n is replaced with $\sqrt{}$ p′ (1−p′) / (n+4).

It is an adjustment comparable to that of the degrees of freedom adjustment of the t-table, but much more easy. What we test is, whether the z-value = the ratio d/SE is larger than around 2 (1.96 for proportions, and 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: $\text{mean}_1 - \text{mean}_2 = 1.4$.

Then calculate pooled SEM $= \sqrt{\left(SEM_1^2 + SEM_2^2\right)} = 0.930$.

Note that, for SEM of difference, take the square root of sums of squares of separate SEMs and, so, reduce the analysis of two means to one of a single mean.

$T = \frac{\text{mean}_1 - \text{mean}_2}{\text{Pooled SEM}} = 1.4/0.930 = 1.505$, with 18 degrees of freedom (dfs)

p>0.05. We have 2 groups of n=10 which means 2×10−2 = 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

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

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

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

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

Note that the z-test uses the bottom row of the traditional t-table, 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.

2.2 Method 2, Chi-Square Test

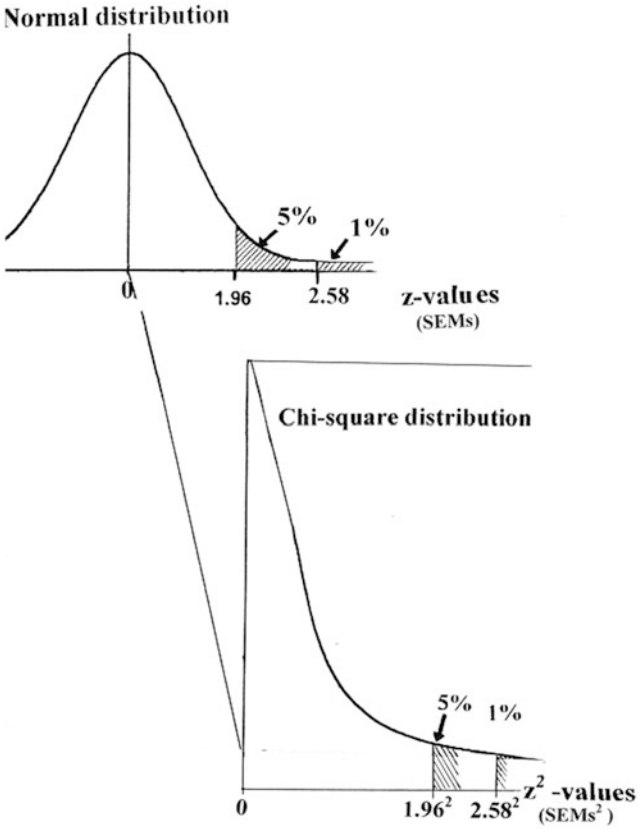
According to some a more easy way to analyze proportional data is the chi-square test. The chi-square test assumes that the data follow a chi-square frequency distribution which can be considered the square of a normal distribution. First some philosophical considerations.

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 (underneath figure, 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 $(\text{normal frequency distributions})^2$ (underneath figure, 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.



The chi-square test, otherwise called χ^2 test can be used for the analysis of two unpaired proportions (2×2 table), but first we give a simpler example , a 1×2 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 " " $= \underline{5 +}$
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).

This 1×2 table has 1 degree of freedom. The chi-square table 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 1×2 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) × $\text{number}_{\text{group}}$).

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 1×2 and the 2×2 contingency tables have only 1 degree of freedom.

2.3 Method 3, Pocket Calculator Method

Instead of the above calculations to find the chi-square value for a 2×2 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.

2.4 Method 4, Fisher Method

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 5×4×3×2×1.

	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 for Analyzing More than Two Unpaired Proportions

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 : ..

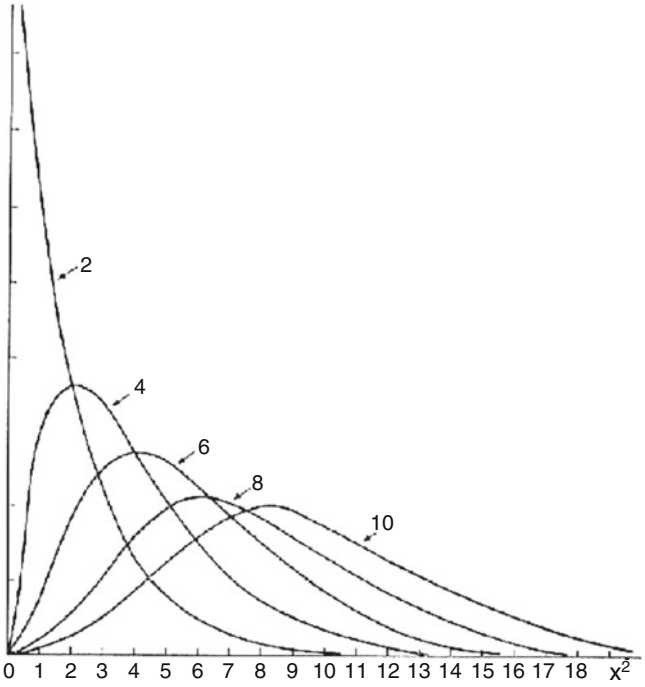
chi-square value = .. +

For cell a $O = 5$

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

Large tables have many degrees of freedom (dfs). For 2×2 cells, we have (2−1) × (2−1) = 1df, 5% p-value at chi-square = 3.841. For 3×2 cells, we have (3−1) × (2−1) = 2dfs, 5% p-value at chi-square = 5.991. For 5×2 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 below):

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.



As an example we give a χ^2 test for 3×2 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 2×2 contingency table above. Per cell: divide hypertensives in study by observations in study, multiply by observations in group. It gives you the best fit estimate. For cell a this is $\alpha = [(a+b+c) / (a+b+c+d+e+f)] \times (a+d)$. 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 $(3-1) \times (2-1) = 3$ degrees of freedom is <0.001 according to the chi-square table.

Another example is given, a 2 × 3 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.

α = [(a+b) / (a+b+c+d+e+f)] x (a+c+e) = 55.000
β .... = 55.000
γ = [(c+d) / (a+b+c+d+e+f)] x (a+c+e) = 51.613
δ = ... = 51.613
ε ... = 55
ξ ... = 55

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

$\chi^2 = 4.01$

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

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 underneath).

		Test 1 +	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). The table 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 (next section).

5 Multiple Paired Binary Data (Cochran's Q Test)

The scientific question of the underneath data is: is there a significant difference between the numbers of responders who have been treated differently three times. Responders (1) and non-responders (0) after treated differently for three times are underneath.

Variables														
1	2	3	1	2	3	1	2	3	1	2	3
0	0	0	0	1	0	0	1	0	1	0	0
0	0	1	0	0	1	0	0	0	1	0	0
0	0	0	0	1	1	0	0	0	1	0	1
0	0	1	0	1	0	0	1	1	1	0	1
0	0	1	0	0	0	0	0	1	1	1	0
0	0	1	0	1	0	0	1	1	1	1	1
0	0	1	0	0	1	0	0	1	1	1	0
0	0	0	0	1	1	0	0	0	1	1	1
0	1	0	0	0	1	0	0	0	1	0	0
0	1	1	0	1	0	0	1	0	1	0	1
0	1	1	0	1	0	0	1	0	1	0	0
0	0	1	0	0	0	0	1	1	1	0	1
0	1	1	0	0	1	0	0	0	1	1	1
0	0	1	0	0	1	0	0	1	1	0	1
0	1	0	0	1	0	0	0	0	1	1	0
0	0	0	0	1	1	0	1	1	1	0	1
0	1	0	0	1	0	0	0	0	1	1	0
0	0	1	0	0	1	0	1	1	1	1	1
0	0	1	0	0	0	0	0	0	1	0	0
0	0	1	0	1	1	0	0	1	1	0	1
0	0	1	0	1	0	0	1	0	1	1	0
0	1	0	0	0	0	0	0	0	1	1	1
0	0	0	0	1	1	0	1	0	1	0	0
0	1	0	0	0	1	0	0	1	1	0	0
0	0	0	0	1	0	0	1	1	1	1	0
0	0	1	0	0	0	1	0	1	1	1	0
0	0	1	0	0	0	1	0	1	1	0	1
0	0	1	0	0	1	1	1	0	1	0	0
0	1	1	0	1	1	1	1	0	1	1	1
0	1	1	0	1	1	1	1	1	1	1	0
0	1	0	0	0	0	1	0	1	1	0	1
0	1	0	0	0	0	1	1	0	1	0	0
0	0	1	0	1	1	1	0	0	1	1	1
0	0	1	0	1	1	1	1	1	1	1	1
0	0	0	0	1	1	1	1	1	1	0	0
									1	0	0

Var 1 = responder to treatment 1 (yes or no, 1 or 0) (Var = variable)
Var 2 = responder to treatment 2
Var 3 = responder to treatment 3

The above table shows three paired observations in one patient. The paired property of these observations has to be taken into account because of the, generally, positive correlation between paired observations. Cochran's Q test is appropriate for that purpose.

For the analysis of these using the Cochran's Q test the following commands have to be given in SPSS statistical software.

Command Analyze....Nonparametric Tests....K Related Samples....mark: Cochran's Q....test variables: enter treatment 1, treatment 2, treatment 3....click OK.

Test Statistics

N	139
Cochran's Q	10,133[a]
df	2
Asymp. Sig.	,006

a. 0 is treated as a success.

The above sheet is in the output sheets. The Cochran test is highly significant with a p-value of 0.006. This means that there is a significant difference between the treatment responses. However, we do not know where: between treatments 1 and 2, 2 and 3, or between 1 and 3. For that purpose three separate McNemar's tests have to be carried out.

Test Statistics[b]

	treat 1 & treat 2
N	139
Chi-Square[a]	4,379
Asymp. Sig.	,036

a. Continuity Corrected

b. McNemar Test

Test Statistics[b]

	treat 1 & treat 3
N	139
Chi-Square[a]	8,681
Asymp. Sig.	,003

a. Continuity Corrected

b. McNemar Test

Test Statistics[b]

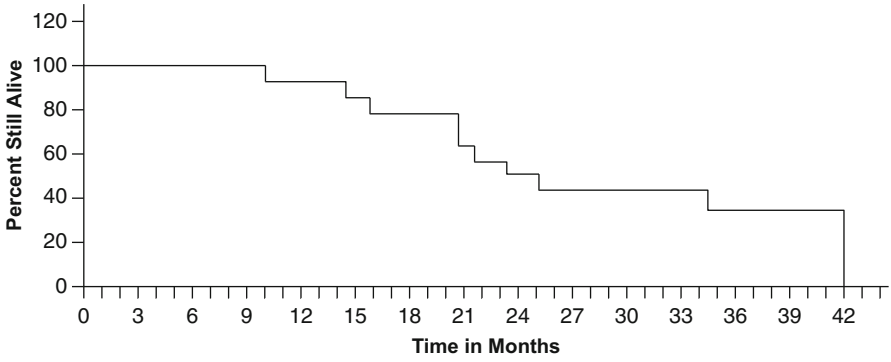
	treat 2 & treat 3
N	139
Chi-Square[a]	,681
Asymp. Sig.	,409

a. Continuity Corrected

b. McNemar Test

The above three separate McNemar's tests show, that there is no difference between the treatments 2 and 3, but there are significant differences between 1 and 2, and 1 and 3. If we adjust the data for multiple testing, for example, by using $p = 0.01$ instead of $p = 0.05$ for rejecting the null-hypothesis, then the difference between 1 and 2 loses its significance, but the difference between treatment 1 and 3 remains statistically significant.

6 Survival Analysis



Above an example of a survival curve plotting survival as a function of time is given.

Mortality with Kaplan-Meier (KM) analysis is an adverse effect in many trials, and KM mortality analysis is pretty much similar to the assessment of proportional data. However, with Kaplan-Meier plots, instead of a single comparison of treatment data versus control, multiple comparisons are assessed and simply added up.

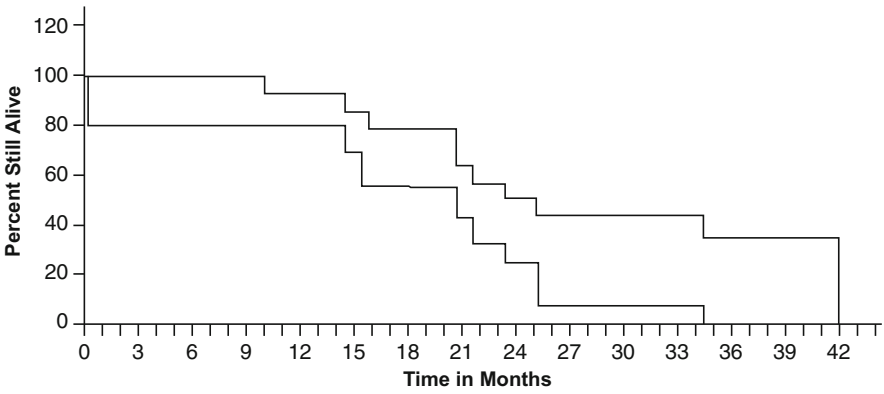
A survival curve plots percentage survival as a function of time. Figure 4 is an example. Fifteen patients are followed for 36 months. At time zero everybody is alive. At the end 40% (6/15) patients are still alive. Percentage decreased whenever a patient died. A problem with survival analysis generally 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 equation:

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. More important than the confidence intervals are tests of significance of difference between two Kaplan-Meier 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. 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 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:

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

where we thus have multiple 2 × 2 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 2 × 2-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.

7 Odds Ratio Method for Analyzing Two Unpaired Proportions

Odds ratios increasingly replace chi/square tests for analyzing 2 × 2 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 logarithmic 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

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

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

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

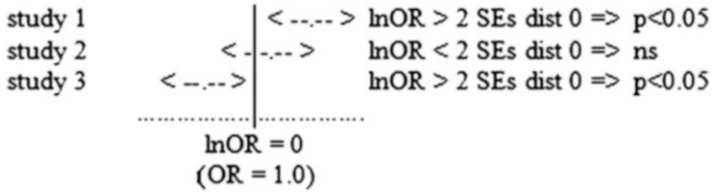
It was astonishing but not unexpected that mathematicians discovered that frequency distributions of log OR followed a normal distribution, and that results were even better if ln instead of log was used.

	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:



Using this method we can test the OR. However, we need to know how to find the SE of our OR. SE of lnOR is given by the formula $\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}$.

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

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

$\text{lnOR} = -1.3863$

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

$\ln \text{ OR} \pm 2 \text{ SEMs} = -1.3863 \pm 1.5182$

= between − 2.905 and 0.132,

Now turn the ln numbers into real numbers by the antiln button of your pocket calculator (2ndf and ln button in many pocket calculators).

=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

$\text{OR} = \frac{103/46}{77/62} = \frac{2.239}{1.242} = 1.803$
$\text{lnOR} = 0.589$
$\text{SEM lnOR} = \sqrt{\left(\frac{1}{103} + \frac{1}{46} + \frac{1}{77} + \frac{1}{62}\right)} = 0.245$
$\text{lnOR} \pm 2 \text{ SEMs} = 0.589 \pm 2(0.245)$
$= 0.589 \pm 0.482$
= 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.

8 Odds Ratios (OR)s for One 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 with drug 2	yes	(a) 65	(b) 28
	no	(c) 12	(d) 34

Here the OR = b/c, and the SE is not $\sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}$, but rather $\sqrt{\left(\frac{1}{b} + \frac{1}{c}\right)}$.
$\text{OR} = 28/12 = 2.33$
$\text{lnOR} = \ln 2.33 = 0.847$
$\text{SE} = \sqrt{\left(\frac{1}{b} + \frac{1}{c}\right)} = 0.345$
$\text{lnOR} \pm 2 \text{ SE} = 0.847 \pm 0.690$
= 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.

9 Loglikelihood Ratios

Loglikelihood ratios are used for safety data analysis. 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, although infrequently used so far.

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 × 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:

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

$$n! = n \text{ faculty} = n\,(n-1)(n-2)(n-3)\ldots\ldots\ldots$$

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

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

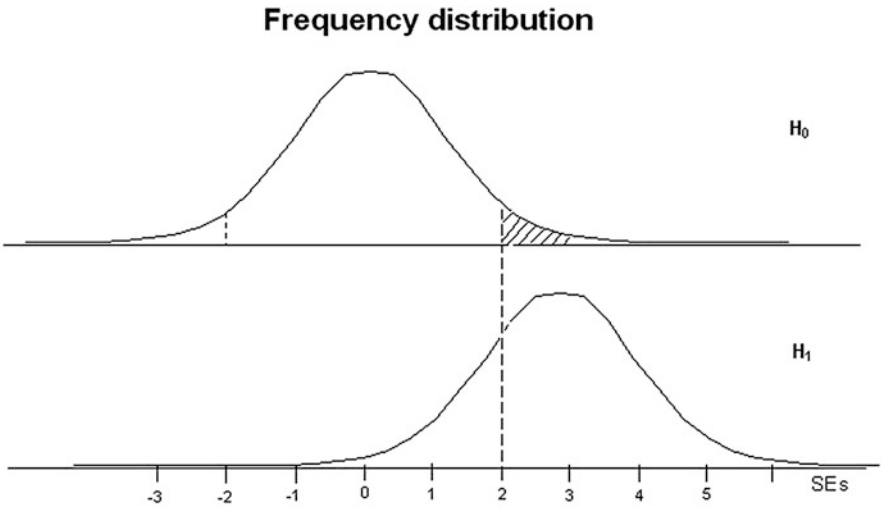
Likelihood p_{ci} can be calculated similarly.

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. Log means natural logarithm, otherwise called naperian logarithm, otherwise called logarithm to the base e.

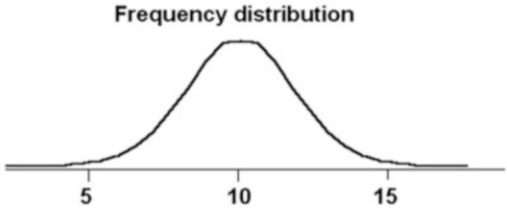
9.1 The Normal Approximation and the Analysis of Events

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.



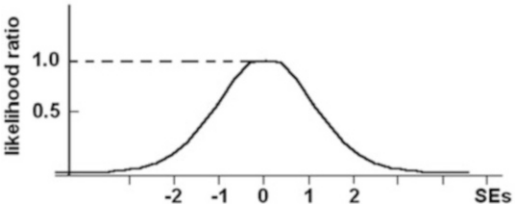
The above H1 = graph based on the data of a sample with standard errors distant from zero (SEs) as unit on the x-axis, often called z-axis in statistics. H0 = same graph with a mean value of 0. We make a giant leap from the sample to the entire population, and we can do so because the sample is assumed 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.



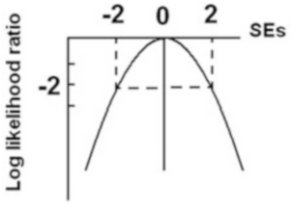
In the above graph, assume that, on average, 10 of 15 patients in a population will have some kind of cardiovascular event within a certain period of time. Then, 10 / 15 will be the proportion most frequently encountered when taking many random samples of 15 patients from this population. The chance 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.

9.2 *Loglikelihood Ratio Tests and the Quadratic Approximation*



In the above graph, assume 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 the graph can also be interpreted as a likelihood ratio curve.



In the above graph we have transformed the likelihood ratio values of the y-axisto log likelihood ratio values, leaving the z – axis unchanged: a different type of 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, the graph 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 is adequate.

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

If we transform the likelihood ratio values of the y-axis to log likelihood ratio values, leaving the z-axis unchanged, then the next equations and their corresponding curve 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 underneath binomial equation.

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

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

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

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

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

$$\begin{aligned}\text{Log likelihood ratio} &= 4 \log \frac{1/310000}{4/300293} + 300289 \log \frac{1-1/310000}{1-4/300293} \\ &= -2.641199 \\ -2 \text{ log likelihood ratio} &= 5.2824 (p < 0.05, \text{because } z > 2).\end{aligned}$$

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

9.3 More Examples

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.

$$\text{Log likelihood ratio} = 9 \log \frac{5/15}{9/15} + 6 \log \frac{1 - 5/15}{1 - 9/15}$$
$$= -2.25$$
$$-2 \log \text{likelihood ratio} = 4.50 \quad (p < 0.05, \text{because } z > 2).$$

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

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.

$$\text{Log likelihood ratio} = 103 \log \frac{77/139}{103/149} + 46 \log \frac{1 - 77/139}{1 - 103/149}$$
$$= -5.882$$
$$-2 \log \text{likelihood ratio} = 11.766 (p < 0.002, \text{because } z > 3.090).$$

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

Traditional statistical tests for the analysis of clinical 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

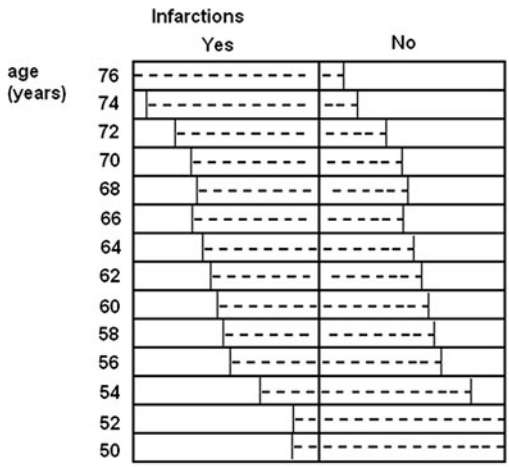
conveniently handled by the log transformation and an exponential equation is turned into a simpler quadratic equation. A potential disadvantage of numerical problems is avoided by taking in the final analysis the ratios of likelihoods instead of separate likelihoods. Log likelihood ratio tests are consistently more sensitive than traditional statistical tests.

10 Logistic Models

Logistic regressions are used for predicting the probability of an event. The odds of an infarction is given by the equation

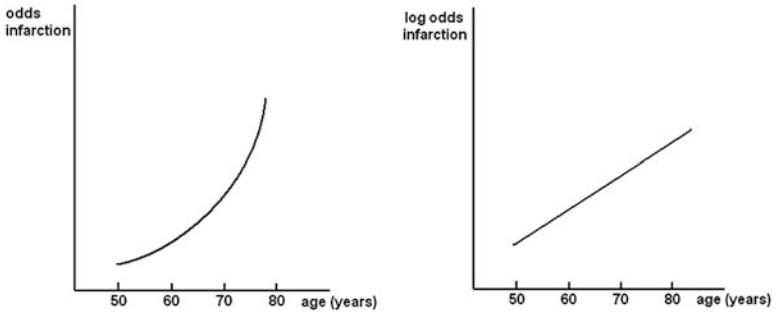
$$\text{odds infarct in a group} = \frac{\text{number of patients with infarct}}{\text{number of patients without}}$$

The odds of an infarction in a group is correlated with age, the older the patient the larger the odds



Number patients at risk with infarct yes / no

In the above graph, in a group of multiple ages the numbers of patients at risk of infarction is given by the dotted line. According to the graph the odds of infarction is correlated with age, but we may ask how?



In the above graphs the relationships between the odds of infarction and age are given. According to the graphs 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:

$$\text{Ln odds55 years} = -9.2 + 0.1.55 = -4.82265$$
$$\text{odds} = 0.008 = 8/1000$$
$$\text{Ln odds75 years} = -9.2 + 0.1.75 = -1.3635$$
$$\text{odds} = 0.256 = 256/1000$$

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

$$\text{“ ln odds infarct} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots.\text{”}$$

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

$$\text{risk event} = 1/(1 + 1/\text{odds})$$

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. 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.

11 Poisson Regression

Poisson regression cannot only be used for counted events per person per period of time, but also for numbers of yes/no events per population per period of time. It is, then, similar to logistic regression, but also different from it, in that it uses a log instead of logit (log odds) transformed dependent variable. It is more adequate and often provides better statistics than logistic regression does, because, again, unlike with logistic regression, time is explicitly included.

Explicit time dependent methods calculate the state of a system at a later time from the state of the system at the current time, while implicit methods find a solution by solving an equation involving both the current state of the system and the later one. Mathematically, if y(t) is the current system state and y(t + Δ t) is the state at the later time, where Δ t is a small time step, then, for an explicit method the underneath equation is adequate

$$y(t + \Delta t) = F(y(t)).$$

An data example of the effect of two parallel-group treatment modalities on the presence or not of a severe cardiac arrhythmia (torsade de pointes) is used. The data file is in extras.springer.com and is entitled "poisson". The first 10 patients are in the underneath table.

treat	presence of torsade de pointes.
,00	1,00
,00	1,00
,00	1,00
,00	1,00
,00	1,00
,00	1,00
,00	1,00
,00	1,00
,00	1,00
,00	1,00

SPSS statistical software will be used for analysis. Start with opening the data file in your computer that has SPSS installed. We will start with a traditional binary logistic regression analysis. Then command.

Command Analyze. . . .Regression. . . .Binary Logistic. . . .Dependent: torsade. . . . Covariates: treatment. . . .click OK.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VAR00001	1,224	,626	3,819	1	,051	3,400
	Constant	-,125	,354	,125	1	,724	,882

a. Variable(s) entered on step 1: VAR00001.

The above table shows that the treatment is not statistically significant. A Poisson regression will be performed subsequently. For this analysis the module Generalized Linear Models is required. It consists of two submodules: Generalized Linear Models and Generalized Estimation Models.

Command Analyze....Generalized Linear Models....Generalized Linear Models . . . .mark Custom. . . .Distribution: Poisson . . . .Link Function: Log. . . .Response: Dependent Variable: torsade. . . . Predictors: Factors: treat....click Model....click Main Effect: enter "treat. . . ..click Estimation: mark Robust Tests. . . .click OK.

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-,288	,1291	-,541	-,035	4,966	1	,026
[VAR00001=,00]	-,470	,2282	-,917	-,023	4,241	1	,039
[VAR00001=1,00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: torsade
Model: (Intercept), VAR00001

a. Set to zero because this parameter is redundant.

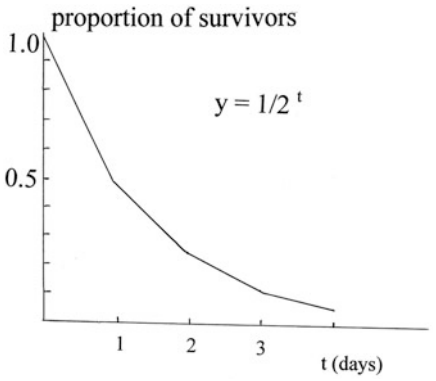
b. Fixed at the displayed value.

The above table is in the output. It shows the results of the Poisson regression for binary outcomes. The predictor treatment modality is statistically significant at $p = 0.039$. According to the Poisson model the treatment modality is a significant predictor of torsades de pointes.

12 Cox Models

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. The underneath graph shows, that. after 1 day, 50% is alive, while, after the second day, 25% is, etc.



And so, an exponential surviving pattern of mosquitos is observed. The formula for the proportion of survivors is given by:

$$\text{proportion survivors} = 1/2^{t} = 2^{-t}$$

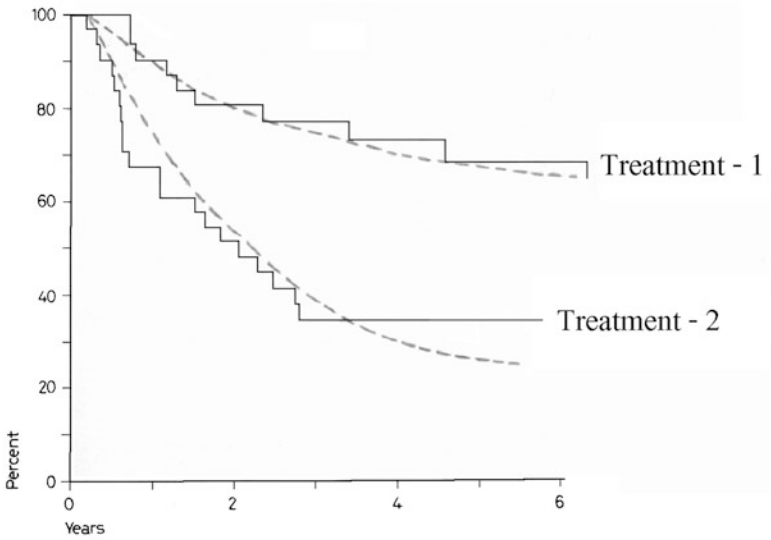
In true biology "e (= 2.71828)" instead of "2" better fits the observed data, while k is dependent on the species:

$$\text{proportion survivors} = e^{-kt}$$

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 $= e^{b}$.

The underneath graph 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.



The above 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 $= e^{b_1 x_1 + b_2 x_2 + b_3 x_3}$
$x_1 = 0$ (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. The figure 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 that Cox regression can be used for other exponential time relationships like pharmacokinetic data. Limitations similar to ones described above apply to such analyses.

13 Bayesian Crosstabs

Classical statistics uses the scientific method to assess whether a new treatment works or not, and it does so by trying and rejecting the null hypothesis of no effect. The classical definition of the null hypothesis is the summary of the means of many trials similar to our trial with an overall effect of zero. The scientific method is considered a gem and the "non plus ultra" for scientific research, because it is less biased than observational data, particularly if applied in prospective randomized controlled trials. However, it will only work, if you accept a number of untested assumptions, for example, that your study sample is representative for the entire

population, and that we have numerous repetitions of trials and that all of them have virtually identical means and standard deviations while, in reality, we have only a single trial. Pretty strong untested assumptions include the study is representative, meaning that, if we repeat it, differences will be small, and all similar studies will have the same standard deviation or error. With Bayesian statistics, there is no null hypothesis, like there is with classical statistics. Yet it can be used to estimate, whether a new treatment works or not. Estimations are performed with likelihood distributions, rather than normal distributions. Why may a likelihood distribution be better than a Gaussian-like normal distribution. That is, because it runs from 0 to ∞, while the latter runs from 0 to 1 (100%). Therefore, it is mathematically better assessable, and can be added, subtracted, divided and multiplied, and even integrated and differentiated. SPSS statistical software version 25 (2017) has started to provide a combined module entitled Bayesian Statistics including almost all of the modern Bayesian tests (Bayesian t-tests, analysis of variance (anova), linear regression, crosstabs etc.).

In studies with both a binary outcome (for example event yes/no) and binary predictor variable (for example treatment group 1 or 2) for traditional analysis a 2×2 interaction matrix is drawn with the predictor in two rows and the outcome in two columns. The null hypothesis H0 is defined as "no difference between the treatment groups", the alternative hypothesis H1 is defined as "a real difference between the treatment groups". Usually the chi-square test is applied for assessing whether the null hypothesis of no difference between the two groups can be rejected. If not, the alternative hypothesis can not be accepted because this was not assessed. Instead of a 2×2 chi-square test also a Bayesian loglinear regression is possible. It assesses the magnitude of the Bayes factor (BF). A BF smaller than "one" supports the above alternative hypothesis (H1), while a BF larger than "one" supports the above null hypothesis (H0). The BF is computed as the ratio of two likelihood distributions, that of the posterior and the prior likelihood distribution. The posterior is modeled from the measured proportions of patients with an event, the prior is modeled with the help of a conjugate prior, which is a way for computing a prior the same as that of the measured posterior likelihood distribution, however, with a standard error of "1".

The computation of the BF requires integrations for accuracy purposes. But, then, it can be used as a statistical index to pretty precisely quantify the amount of support in favor H1 and H0. Advantages of the Bayesian approach may include.

1. A better underlying structure model of the H1 and H0 may be provided.
2. Maximal likelihoods of likelihood distributions are not always identical to the mean effect of traditional tests, and this may be fine, because biological likelihoods may better fit biological questions than numerical means of non-representative subgroups do.

However, in spite of this, nobody knows for sure why likelihood distributions may better than normal distributions estimate uncertainties in statistical test results. So, why not use both of them for analyzing the same data example. Underneath the results of traditional 2×2 chi-square-tests and the Bayesian loglinear regression for 2×2 interaction matrices will be shown and compared. For self-assessment a data file is in extras.springer.com, and is entitled "bayesian".

The underneath table are the data of a study of 30 patients admitted to two hospital departments, surgery and internal medicine. The primary scientific question was: is there a significant difference between the risks of falling out of bed at the departments of surgery and internal medicine.

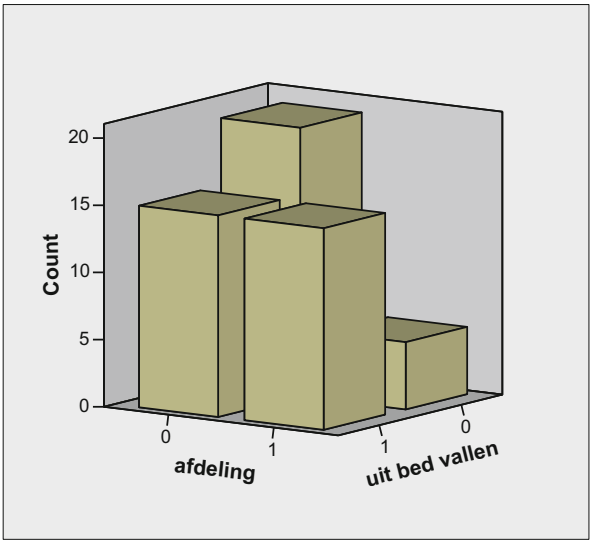
	Fall out of bed	
	no	yes
Number of patient department surgery (0)	20	15
internal department (1)	5	15

The above interaction matrix of the data shows that at both departments the same numbers of patients fall out of bed. However, at the department of surgery many more patients do not fall out of bed than at the internal department.

13.1 Traditional Analysis for 2×2 Interaction Matrix

The data file stored at extras.springer.com and entitled "bayesian" will first be used for drawing a three dimensional graph of the data. Open the data in SPSS statistical software version 25 (with the Advanced Statistics module included) by clicking the title in your computer mounted with SPSS. Then command.

Command click Graphs....click 3D Charts....X-Axis: enter departments....Z-Axis: enter falling out of bed....click OK.



At both departments approximately the same number of patients fall out of bed. However, at department-0 many more patients do not fall out of bed than at department-1 do. The traditional analysis for a 2×2 interaction matrix is tested below.

Command Analyze....Descriptive Statistics....Crosstabs....Rows: enter variable 1....Columns: enter variable 2....Statistics....Chi-Square....click OK.

The table below is in the SPSS output sheets.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,304[b]	1	,021		
Continuity Correction[a]	4,086	1	,043		
Likelihood Ratio	5,494	1	,019		
Fisher's Exact Test				,027	,021
Linear-by-Linear Association	5,207	1	,022		
N of Valid Cases	55				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,09.

The chi-square test (Pearson Chi-Square) shows, that a significant difference between the surgical and internal departments exists in patterns of patients falling out of bed. The p-value equals 0.021, and this is much smaller than 0.05. Several contrast tests have been given in the above results table. They produce approximately similar p-values. This supports the accuracy of the chi-square test for these data.

We have cells with only 5 counts, and, so, the traditional chi-square is flawed. We can nonetheless conclude that the null hypothesis of no difference between the two departments can be rejected. We can not conclude from this, that the alternative hypothesis, a true difference between the two departments, is true, because this was not assessed. In order to better assess the real meaning of our data result, and answer the question what level of likelihood we have to support either the null (H0) or the alternative hypothesis (H1), a modern Bayesian linear regression analysis will be performed. For that purpose in SPSS statistical software module version 25 a Bayesian loglinear regression analysis will be performed.

13.2 Bayesian Loglinear Regression for 2 × 2 Interaction Matrix

Command Analyze...Bayesian Statistics....Loglinear Models....Row variable: enter department ....Column variable: enter outcome....Bayesian Analysis: mark Use Both Methods....click Criteria: leave default setting....click Continue....click Bayes Factor: mark Conjugate, otherwise default setting....click Continue.... click OK.

The four tables below are in the output sheets.

Case Processing Summary

	Valid		Missing		Total	
Observed Cases	N	Percent	N	Percent	N	Percent
department * outcome	55	90,2%	6	9,8%	61	100,0%

department * outcome Category Tabulation

			outcome		
			,00	1,00	Total
department	,00	Count	20	15	35
	1,00	Count	5	15	20
Total		Count	25	30	55

Test of Independence [a]

	Value	df	Asymptotic Sig.(2-sided)	Exact Sig.(2-sided)	Exact Sig.(1-sided)
Bayes Factor	,161 [b]				
Pearson Chi-Square	5,304 [c]	1	,021		
Continuity Correction	4,086	1	,043		
Fisher's Exact Test				,027	,021

a. The total sum is fixed in the contingency table.

b. This analysis tests independence versus association, and assumes a multinomial model and conjugate priors.

c. 0 cells(0,0%) have expected count less than 5. The minimum expected count is 9,091.

Posterior Distribution Characterization of Simulated Interactions[a,b]

	Posterior			95% Simultaneous Credible Interval		
Interaction	Median	Mean	Variance	Lower Bound	Upper Bound	Contains 0 or not
,00, ,00	1,301	1,314	,359	,176	2,528	No

a. The analyses assume an independent multinomial model.

b. Seed: 599767049. Number of simulated posterior samples: 1000000.

The first and second Bayesian tables give simply summary statistics. The third table reports the tests of independence versus association, or support for real difference between the departments or not. In addition to the test statistics similar to those of the traditional chi-square tests, the Bayes factor (BF) is given. It is computed from the division sum of posterior and prior likelihood distributions where posterior and prior are based on respectively the observed data and, in the absence of an informed prior, a conjugate measure obtained from the posterior, however with a standard error of 1. The above fourth table gives the statistics of the posterior. Footnote c shows that Monte Carlo simulations were applied to integrate out the nuisance parameter, a parameter not directly of interest to the ratio of likelihood distributions. After many iterations it is zero. The median and mean of the selected posterior distribution are not normal and slightly skewed to the right.

Nonetheless it performed well enough, and the ratio of posterior and prior, the Bayes factor, was pretty small, 0.161. This indicates that we have moderate to strong support for the alternative hypothesis, a real difference between the two departments. This result is in agreement with the above results from the traditional chi-square test.

The underneath Bayes factor table given by SPSS is helpful here.

Reject H1

Bayes Factor	Evidence Category	Bayes Factor	Evidence Category	Bayes Factor	Evidence Category
>100	Extreme Evidence for H0	1-3	Anecdotal Evidence for H0	1/30-1/10	Strong Evidence for H1
30-100	Very Strong Evidence for H0	1	No Evidence	1/100-1/30	Very Strong Evidence for H1
10-30	Strong Evidence for H0	1/3-1	Anecdotal Evidence for H1	1/100	Extreme Evidence for H1
3-10	Moderate Evidence for H0	1/10-1/3	Moderate Evidence for H1		

Reject H0

Bayesian loglinear also produces statistics of counts observed and expected. For the computation the default command is: first click Print. Then mark Expected Counts, Percentages Row, Column, and Total. The underneath table is in the SPSS output.

department * outcome Category Tabulation

			outcome		
			,00	1,00	Total
department	,00	Count	20	15	35
		Expected Count	15,9	19,1	35,0
		% within department	57,1%	42,9%	100,0%
		% within outcome	80,0%	50,0%	63,6%
		% of Total	36,4%	27,3%	63,6%
	1,00	Count	5	15	20
		Expected Count	9,1	10,9	20,0
		% within department	25,0%	75,0%	100,0%
		% within outcome	20,0%	50,0%	36,4%
		% of Total	9,1%	27,3%	36,4%
Total		Count	25	30	55
		Expected Count	25,0	30,0	55,0
		% within department	45,5%	54,5%	100,0%
		% within outcome	100,0%	100,0%	100,0%
		% of Total	45,5%	54,5%	100,0%

Obviously, the departments (0) and (1) are analyzed separately: the proportions "fall out of bed" "yes" or "no" per department are reported both separately and together.

The above Bayesian approach pretty much matches traditional analysis methods.

With modern Bayesian statistics likelihood distributions rather than probability distributions are modeled. It is mostly based on the Cauchy distribution, a member of the family of the alpha distributions, and the distribution that best describes the ratio of either two normal distributions or two likelihood distributions. Bayes factors have a Cauchy distribution, and cannot be numerically analyzed with standard Gaussian approximation methods. Fortunately, pretty good numerical results are obtained by integrations that integrate out nuisance variables.

Advantages of the Bayesian approach may include.

1. A better underlying structure of the H1 and H0 may be provided.
2. Bayesian tests work with 95% credible intervals that are usually somewhat wider and this may reduce the chance of statistical significances with little clinical relevance.
3. Maximal likelihoods of likelihood distributions are not always identical to the mean effect of traditional tests, and this may be fine, because biological likelihoods may better fit biological questions than numerical means of non-representative subgroups do.
4. Bayes uses ratios of likelihood distributions rather than ratios of Gaussian distributions, which are notorious for ill data fit.
5. Bayesian integral computations are very advanced, and, therefore, give optimal precisions of complex functions, and better so than traditional multiple mean calculations of non representative subsamples do.
6. With Bayesian testing type I and II errors need not being taken into account.

A disadvantage of Bayesian methods may be overfitting. This means that the likelihood distributions may be wider than compatible Gaussian modeling. Bootstraps t-test is based on Monte Carlo resampling from your own data. It is available in SPSS statistical software. In the example given we will compare a bootstraps sampling distribution in SPSS with Bayesian likelihood and traditional Gaussian distributions. Once again the data example from the current chapter is used.

Command Analyze....Regression....Binary Logistic Regression....Dependent: fall out of bed....Covariate(s) Department....click Bootstrap....click Perform bootstrapping....Number Samples enter 1000....click Continue....click OK.

The bootstrap resampling model is in the output sheets. It provides a 95% confidence interval between 0.176 and 3.043.

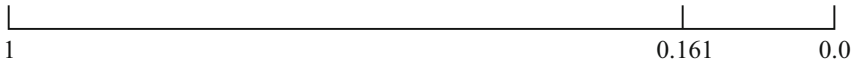
The Bayesian and traditional chi-square 95% confidence intervals are given above.

1. Gaussian 95% confidence interval 0.148 and 2.624
2. Bootstrap 95% confidence interval 0.176 and 3.043
3. Bayesian 95% credible interval 0.176 and 2.528.

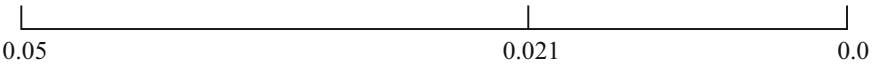
Obviously, the Gaussian confidence interval, the bootstrap confidence interval, and the Bayesian credible interval have very much similarly sized intervals. Overfitting is not obvious.

The Bayes factor (BF) in this chapter's example of 0.161 suggests that the Bayesian crosstab model provides a better sensitivity of testing than the traditional chi-square-test does.

On a continuous line of Bayes factors from 1.0 to 0.0 our Bayes factor is on the right end side.



On a continuous line of p-values from 0.05 to 0.0 our p-value is closer to the middle.



The BF of 0.161 is closer to 0.0 (or very small), than the p-value of 0.021 is. The BF seems to provide a slightly better statistic here than the p-value does.

14 Discussion

1. For the analysis of efficacy data we test null-hypotheses, safety data consist of proportions, and require for statistical assessment different methods.
2. 2 × 2 tables are convenient to test differences between 2 unpaired proportions.
3. Use chi-square or t-test for normal distributions (z-test) for that purpose.
4. A pocket calculator method is demonstrated and the Fisher method is explained.
5. Chi-square tests for analyzing more than two unpaired proportions are shown.
6. For paired proportions the McNemar's test is appropriate.
7. Multiple paired binary data are tested with Cochran's Q test.
8. Survival data are described with Kaplan Meier curves. They are also proportional data, including lost patients in the analysis.
9. Two Kaplan-Meier Curves can be compared using the Mantel-Haenszl = Log rank test
10. Odds ratios with logarithmic transformation provide an alternative method for analyzing unpaired 2 × 2 tables.
11. Odds ratios for one group, two treatments are analyzed with Mc Nemar's odds ratios.
12. Loglikelihood ratios is an exact test for analyzing unpaired 2×2 tables, and it performs better than traditional methods do. This is important because it frequently changes insignificant results into significant ones.
13. Logistic models are another alternative for assessing the probability of an event, just like 2×2 tables do, but because it is a regression model, it can include covariate variables as additional predictors in additin to the treatment modality.
14. Poisson regression cannot only be used for counted events per person per period of time, but also for numbers of yes/no events per population per period of time.

15. Cox models uses exponential curves for testing survival from one treatment versus another. As an regression it just like logistic regression can include covariate for adjustment of the overall results.
16. Bayesian crosstabs is the standard Bayesian method for analyzing interaction matrices. As compared to traditional unpaired chi-squares and log likelihoods, better sensitivity of testing is often obtained.

15 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 3
Incidence Ratios, Reporting Ratios, and Safety Signals Instead of Adverse Effects

Abstract Proportional reporting ratios are statistics that compare (out of all patients with adverse effects) the proportion of patients with a specific adverse effect from a particular drug with that of the same in a much larger group of patients (using the same drug class).

Standardized Incidence Ratios are age-adjusted rates of adverse effects compared to similarly adjusted rates from larger populations like, cities, states, entire countries.

A safety signal is data information, suggesting a new causal association between a medicine and an adverse effect. In the past few years pharmacologists have worldwide been developing criteria for estimating the seriousness of adverse effects, other than or additional to the traditional statistical test criteria as reviewed in the Chap. 2. A problem with safety signals is their incomplete character.

Keywords Proportional reporting ratios · Standardized reporting ratios · Large Chi-Square tables · Safety signals · Pharmacovigilance · Spontaneous reporting systems · Potential signals · Signal detection

1 Introduction

In controlled studies the comparison to a matched control group may be state of the art, but controlled trials are expensive and population data, like data from a city, state, or even an entire country as control, if available, provides more stable statistics. There are of course problems with historical controls, because of the risk of asymmetries due to different times, populations, equipments, but they can be given appropriate notice. For example in a controlled study the proportional reporting ratio can be used (Evans et al., Pharmacoepidemiol Drug Saf 2001; 10: 486), as a statistic that compares out of all patients with adverse effects the proportion of patients with a specific adverse effect from a particular drug compared to that of the same in a much larger group of patients using drugs from a similar drug class.

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_3

Another method is the standardized incidence ratio (Sahai and Khurshid, Biometrical J 1993; 35: 857), where age-adjusted rates of adverse effects are compared to similarly adjusted rates from larger populations, like cities, etc.

2 Chi-Square Test

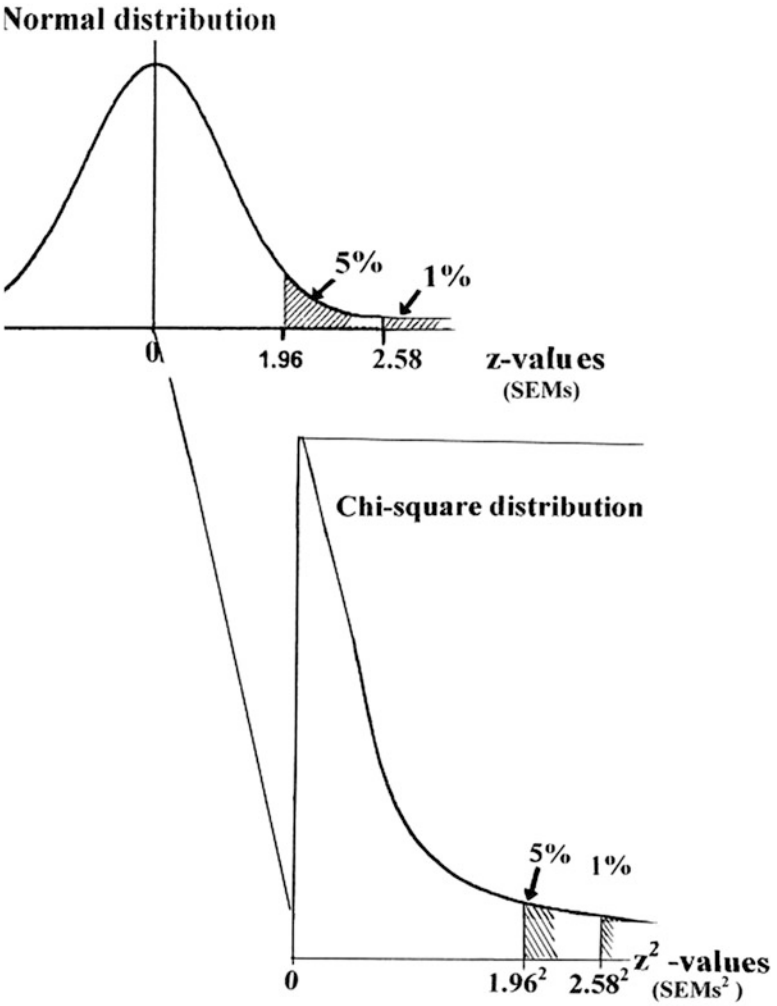
An easy way to analyze proportional data is the chi-square test. The chi-square test assumes that the data follow a chi-square frequency distribution which can be considered the square of a normal distribution. First some philosophical considerations.

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 (underneath figure, 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 $(\text{normal frequency distributions})^2$ (underneath figure, 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.



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

	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

The above chi-square table 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 1×2 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)	10(c)	5 (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: $(O-E)^2 / E = (5 - 15/30 \times 15)^2 / 15/30 \times 15 =$ ..
" b: $(O-E)^2 / E =$ ....
" c: $(O-E)^2 / E$ .....
" d: $(O-E)^2 / E$ ..... +
chi-square = 3.33

(O = observed number; E = expected number = (proportion sleepers /total number) × $\text{number}_{\text{group}}$).

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 3.33 means that the 0-hypothesis can not be rejected. Obviously, the first data example gave a very significant result unlike the second data example. If you have control data from a large population instead of thise from a small control group, you will obtain a statistical model with a lot more power than you will with just a small sized control sample.

Note: a chi-square distribution = a squared normal distribution. When using the chi-square table, both the 1 × 2 and the 2 × 2 contingency tables have only 1 degree of freedom.

3 Proportional Reporting Ratios

Proportional reporting ratios are statistics that compare (out of all patients with adverse effects) the proportion of patients with a specific adverse effect from a particular drug with that of the same in a much larger group of patients (using the same drug class). As an example, in a study 8 of 135 patients had a particular adverse

effect. We could obtain about 4050 patients having had the same class of drug, and having had adverse effects in 160 patients. This meant, that proportions of 8/135 = 0.059 in the small study, and 160/4050 = 0.040 in the large study had adverse effects. This would mean that the proportional reporting ratio is 0.059/0.040 = 1.48. The chance of this adverse effect was about 1.5 times greater in our study than it was in the large data. We need to assess whether this ratio is significantly different from a ratio of 1.0. Chi-square is used.

Instead of the above calculations to find the chi-square value for a 2 × 2 contingency table, the simpler pocket calculator method, already explained in the Chap. 2, can be applied.

	Adverse	no adverse effect	total
small study	8 (a)	127 (b)	a+b
large data	160(c)	3890 (d)	c+d
	a+c	b+d	

Calculating the chi-square (χ^2) – value is done 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 1.28 at 1 degree of freedom which means that the 0-hypothesis of no difference can not be rejected. There is no significant difference between study and control.

4 Standardized Incidence Ratios (SIR)

Another method is the standardized incidence ratio, where age-adjusted rates of adverse effects are compared to similarly adjusted rates from larger populations like, cities, states, entire countries. As an example we will use a slightly modified version of the example of standardized incidence ratios from the School of Public Health University of Boston (sphweb.bumc.bu.edu).

age group	age-adjusted state rate	town size	expected cases
0–19	0.0001	74,850	0.0001 × 74,850
20–44	0.0002	134,950	.....
45–64	0.0005	54,460	....
65–74	0.0015	25,130	...
75–84	0.0018	17,010	..

85+	0.0010	6,330	.
Total			≈136

The add-up sum of the observed cases was 144.

And so, the ratio (observed cases/expected cases) would equal (144/136) × 100–106%.

The bumc website used approximate 95% confidence intervals using the Wilson Hilferty approximation for chi-square percentiles (Math 1931; 17: 685).

Left confidence limit =

$$\text{SIR}\left[1-1/(9\ \text{observed rate}) + Z_{1/2\ \alpha}/(3\sqrt{\text{observed rate}})\right]^3$$

Right confidence limit =

$$\text{SIR}\left[1 - 1/(9\ \text{observed rate}) + Z_{1\ -\ 1/2\ \alpha}/(3\sqrt{\text{observed rate}})\right]^3$$

Left confidence limit =

$$106 \times [1.056]^3 = 124\%$$

Right confidence limit =

$$106 \times [0.944]^3 = 89\%$$

The 95% confidence interval is, thus, between 89% and 124%. Unfortunately, this interval is not significantly different from 100%.

5 Examples of Larger Chi-Square Tables for Comparing the Presence of Adverse Effects Between Different Studies

Example I

As an example we give a χ^2 test for 3 × 2 table. In three studies the presence of hypertension as an adverse effect was assessed. The underneath table summarizes the results. We wished to know, whether the proportion of patients with hypertension were significantly different between the different studies.

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

Give the best estimate of the expected numbers per cell. The estimate is based on the information contained in the summary of the above table. Per cell: divide hypertensives in study by observations in study, multiply by observations in group. It gives

you the best fit estimate. For cell a this is $\alpha = [(a+b+c)/(a+b+c+d+e+f)] \times (a+d)$. Do the same for each cell and add-up:

$$\alpha = [(a+b+c)/(a+b+c+d+e+f)] \times (a+d) \quad = 52.17$$
$$\beta.... \quad = 114.78$$
$$\gamma \quad = 73.04$$
$$\delta = \big[(d+e+f)\big)/(a+b+c+d+e+f)\big] \times (a+d) \quad = 47.83$$
$$\varepsilon.... \quad = 57.39$$
$$\xi.... \quad = 66.96$$

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

The p-value for $(3-1) \times (2-1) = 3$ degrees of freedom is <0.001 according to the chi-square table. This would mean that the three studies are significantly different from one another as far as the numbers of hypertensive patients is concerned.

Example II

Another example is given, a 2 × 3 table. Two studies assess three levels of adverse effects: hypertension (1) present, (2) not present, (3) uncertain.

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

Again 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.

$$\alpha = [(a+b)/(a+b+c+d+e+f)] \times (a+c+e) \quad = 55.000$$
$$\beta.... \quad = 55.000$$
$$\gamma \quad = 51.613$$
$$\delta = [(c+d)/(a+b+c+d+e+f)] \times (a+c+e) \quad = 51.613$$
$$\varepsilon.... \quad = 55$$
$$\xi.... \quad = 55$$

$$
\begin{aligned}
(O\text{-}E)^2 / E &= \\
(a\text{-}\alpha)^2 /\alpha &= 0.45 \\
(b \ldots &= 0.45 \\
(c \ldots &= 0.847 \\
(d \ldots &= 1.363 \\
(e \ldots &= 0.45 \\
(f \ldots &= 0.45 \quad + \\
\chi^2 &= 4.01
\end{aligned}
$$

For $(2 - 1) \times (3 - 1) = 2$ degrees of freedom our p-value is <0.001 according to the chi-square table. This would mean that the two studies are significantly different from one another as far as the levels of hypertension are concerned.

6 Safety Signals Instead of Adverse Effects

Pharmacovigilance is a worldwide movement in clinical pharmacology involved in the practice of monitoring the effects of medicines after they have been licenced for use in order to identify so far unreported adverse effects. In the terminology of pharmacovigilance the terms adverse effect, adverse reaction, adverse event have been largely replaced with the term safety signal or briefly signal. A safety signal is data information, suggesting a new causal association between a medicine and an adverse effect. In the past few years pharmacologists have worldwide been developing criteria for estimating the seriousness of adverse effects, other than or additional to the traditional statistical test criteria as reviewed in the Chap. 2. Puijenbroek (Br J Clin Pharmacol 2001; 52: 579) selected four important factors for a potential adverse effect to be named signal.

That is (1) a new association, (2) a strong association, (3) a serious effect, (4) dechallenge and rechallenge effects. Criteria for the process of renaming a potential adverse effect a signal should include 10 important points (Meyboom, Drug Saf 2002; 25: 459).

1. Unknown adverse reaction.
2. Strong statistical connection.
3. Unexpected.
4. Expected but unlabelled.
5. Specific, characteristic.
6. Objective (definitive) event.
7. Typical drug-related event or critical term.
8. Low background frequency.
9. Serious.
10. High potential relevance.

Obviously, may more criteria than a single statistical test is being checked and approved in the process of signal detection. Nonetheless, in the past 20 years several

national and international pharmacovigilance foundations have been active, and nowadays we have spontaneous reporting systems of potential and definite safety signals. For example, the Uppsala Monitoring Centre has gathered reports from 47 countries of the World Health Organization collaborative program for international drug monitoring consistent of 1.2 million reports. The Dutch pharmacovigilance foundation manages 26,555 reports, while the French pharmacovigilance database (Eudipharm Lyon) manages over 200,000 reports.

Potential signals selected from the reporting systems may be numerous, but only few are according to the managers important enough to become selected signals. As an example from the Dutch pharmacovigilance foundation, out of 26,555 reports between second quarter of 1997 and third quarter of 2000 only 42 were selected as definitive signals. Part of them is given underneath.

pergolide-pulmonary fibrosis	3rd quarter 2000
rofecoxib-death	
clopidrogel-thrombotic thrombocytopenic purpura	
minocycline-interstitial pneumonia	2nd quarter 2000
lamotrigine-Steven Johnson syndrome	
cotrimoxazol-tremor	
loperamide-urinary retention	1st quarter 2000
simvastatin-eczema	
acitretin-taste loss	4th quarter 1999
valproic acid-polycystic ovary syndrome	
metronidazole-hepatitis	
valproic acid-parkinsonism	3rd quarter 1999
lamitrigine-sialoadenitis	
alendronate-alopecia	
interferon-Raynaud	2nd quarter 1999
atorvastatin-rhabdomyolyis	
budesonide-anaphylactic reaction	
diclofenac-haemolytic anemia	
quatiapine-leucopenia	1st quarter 1999
diclofeniac-anaphylactic reaction	
itraconazole-dyspnoea	
sildenafil-death	4th quarter 1998
fexofenadine-QT prolongation	
olanzapine-death	
nefazodone-priapism	
tolcazodone-leucopenia	3rd quarter 1998
vigabatrin-visual field defect	
rulizole-thrombopenia	2nd quarter 1998

One may wonder, if a selection procedure, that includes so many very subjective criteria, has left a scrap of validity. However, the signal selection is, obviously, a learning process, and, at the same time, an ongoing search for better understanding

of contributory factors. Safety signal detection is a field defined as the methodology for identifying information about safety levels of novel food and drug compounds from multiple studies. Those involved are generally very motivated. But the above list can hardly be thought of as complete, and advanced arithmetics of incomplete data may be pretty meaningless. Even so, novel tools are included in the assessments, like relative reporting ratios, Bayesian network meta-analyses. The latter are prone to overdispersion. However, a shrinking method may be used for adjustment, like the gamma Poisson shrinker. Also disproportionality analyses and increasingly newer techniques are being developed to assess safety signals. Computational examples are beyond this introductory edition, but free software packages are available like for example the "openEBGM" (empirical Bayesian geometric mean) package.

7 Discussion

In controlled studies the comparison to a matched control group may be state of the art, but controlled trials are expensive. Population data, like data from a city, state, or even an entire country are increasingly available, and could be used for the purpose of control. These kind of historical control data provides more stable statistics. There are, of course, problems with historical controls, because of the risk of asymmetries due to different times, populations, equipments, but these problems can be given appropriate notice. These kinds of historical control data provide more stable statistics. They provide less spread than the standard errors of small control samples. For example, in a controlled study the proportional reporting ratio can be used, as a statistic, that compares out of all patients with adverse effects the proportion of patients with a specific adverse effect from a particular drug compared to that of the same in a much larger group of patients using the same drug class. Another method is the standardized incidence ratio, where age-adjusted rates of adverse effects are compared with similarly adjusted rates from larger populations, like cities or even entire countries, etc.

8 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,

Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 4
Safety Analysis and the Alternative Hypothesis

Abstract The type I error is the chance of finding a difference where there is none. The type II error is the chance of finding no difference where there is one.

With safety analyses we try and reject the alternative hypothesis of an adverse effect, with efficacy analyses we try and reject the null hypothesis of no treatment effect.

With efficacy analyses the null hypothesis is usually rejected with a type I error of 5% and a type II error may very well be as large as 50%. With safety analysis we are more interested in smaller type II errors, because we will have increasing chance of rightly rejecting the alternative hypothesis, and that is good. After all, the main purpose of safety analyses in clinical trials is to find no adverse effects rather than the opposite. With a type I error of 20% instead of 5% we can reject the type II error at 25%. This means, that the chance of rightly rejecting the alternative hypothesis is 75%, rather than 50%. This result is in better agreement with the incentive to rightly reject the alternative hypothesis, which is the main incentive of safety assessments in clinical trials.

Keywords Safety analyses · Alternative hypothesis · Null hypothesis · Type I error (Alpha) · Type II error (Beta) · Efficacy analysis · Safety analysis · Minimized betas · Flexible alphas and betas

1 Introduction

The alternative hypothesis is traditionally used for power assessments, but it can also be applied for the purpose of safety analysis. In the previous chapter we discussed traditional statistical methods to test the presence of statistically significant adverse effects in clinical trials. This chapter will address modified and improved statistical methods as recommended in the past few years. Particularly the FDA's (American Food and Drug Administration) final rules from March 2011 for expedited reporting

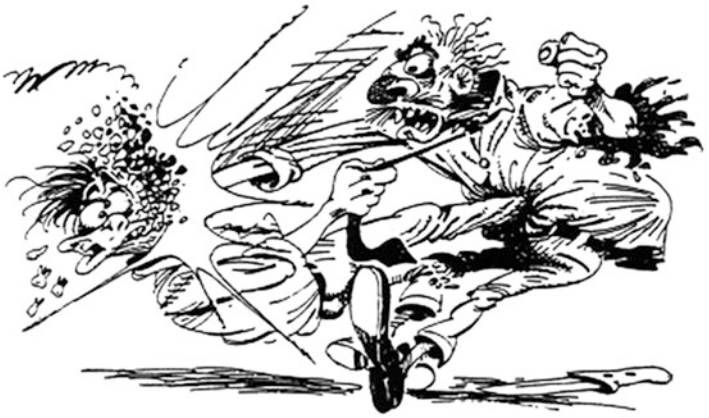
T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_4

of serious adverse events for studies conducted under an investigational New Drug Application gave important modifications for drug safety analysis. It concerned three classes of serious adverse effects. The class A needed not reported, B and C needed expedited report to the FDA. The main purpose of the new rules were

1. less adverse effect reports overall, but
2. more reports correctly identifying them.

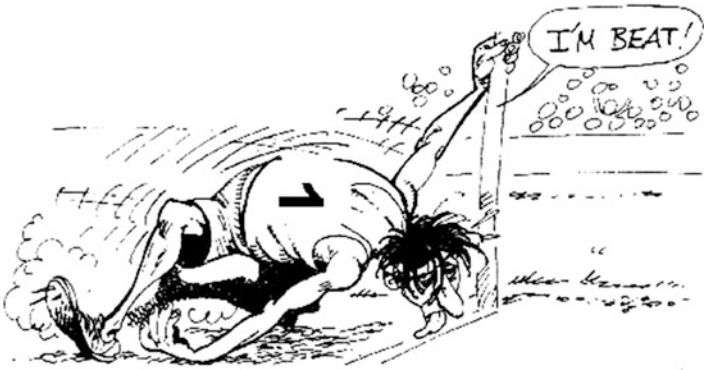
2 Power and the Alternative Hypothesis

Clinical trials test as possible difference between new and standard treatment (or placebo). Statistical power is the chance of finding a difference where there is one.

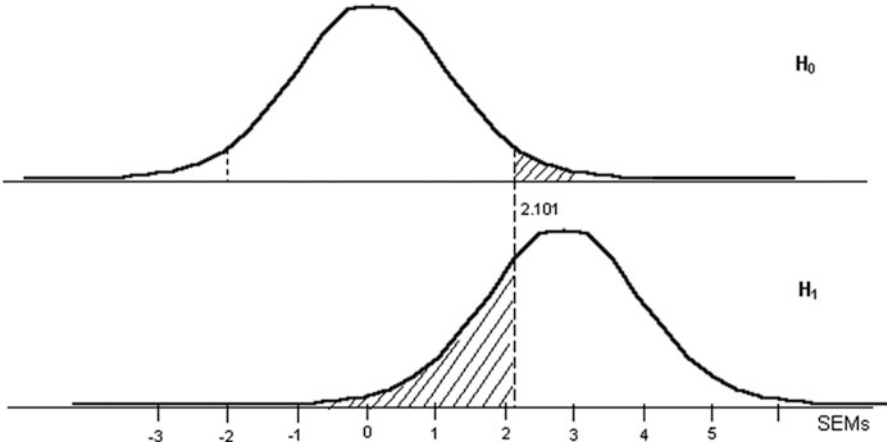


Big power is a big chance of finding a difference where there is one. Large trials have big power. Other relevant possibilities are:

the chance of finding no difference where there is one (type II error)
the chance of finding a difference where there is none (type I error).
Important hypotheses are
Hypothesis 0 (no difference from a zero effect)
Hypothesis 1 (a real difference from a zero effect).



We will now particularly emphasize the hypothesis 1. The underneath graphs are helpful for the purpose.



H1 is the graph based on the data of our trial (mean ± standard error of the mean (SEM)).
H0 is the same graph with mean 0 (mean ± SEM).
H1 is also the summary of the means of many trials similar to ours.
H0 is summary of many trials similar to ours but with overall effect 0.

If hypothesis 0 is true, then the mean of our study will be part of H0.
If hypothesis 1 is true, then the mean of our study will be part of H1.
So, the mean of our study may be part of H0 or of H1.
We can not prove, but calculate the chance of either of these possibilities.

A mean results of 2.9 as observed is far from a result of 0.
Suppose it belongs to H0.
Only 5% of the H0 trials are >2.1 SEMs distant from 0.
The chance, that it belongs to H0 is <5%. We will reject this small possibility.

Suppose the result belongs to H1.
Up to 30% of the H1 trials are <2.1 SEMs distant from 0.
These 30% can not reject the null hypothesis of no effect.
Right from 2.1 SEMS (an area under the curve of 70%) can do so.

Conclusions of the above reasonings:

If H0 is true, we will have <5% chance to find it.
If H1 is true, we will have 70% chance to find it.
This would mean that we can reject the null hypothesis of no effect at $p < 0.05$, and with a power of 70%.

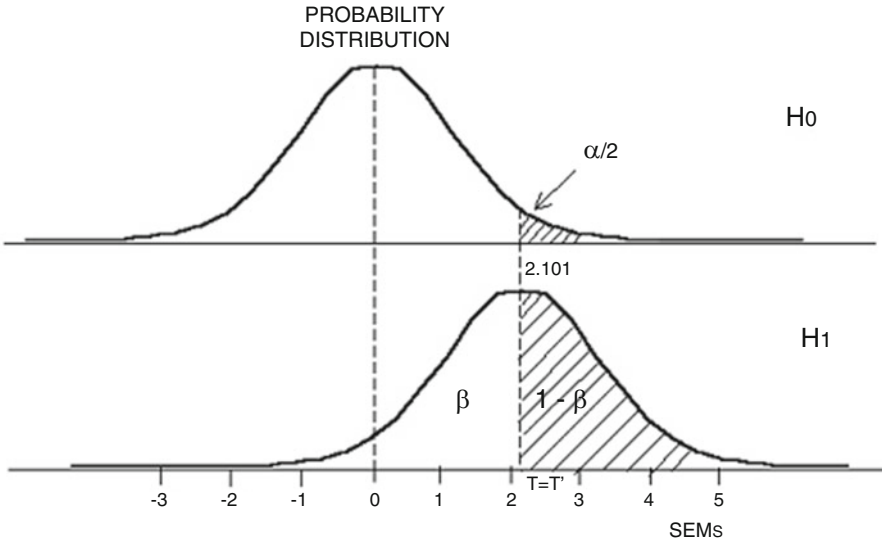
The alternative hypothesis can not only be used for power assessments, but also for the purpose of safety assessment. We should add that the alternative unlike the null hypothesis is always assessed one-sided. We will now address the subject of safety assessment using the alternative hypothesis in particular.

3 Two Main Hypotheses of Clinical Research, Efficacy and Safety

Drug trials are mainly for addressing the efficacy as well as the safety of the treatments 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, or compute confidence intervals to estimate the magnitude of the treatment effect. The appropriate test depends upon the type of data. 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. The usual 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.

However, relative risks are kind of tricky, because they do not take into account the absolute prevalences of events in the treatment and control groups, but only their ratios. For example, if your relative rise of survival rises 30%, your absolute risk reduction may be as little as 1%. This is because, if you go from 3% risk of death to 2%, then the absolute difference is 1%, and the relative difference is 33%. We should add that patients often prefer a better quality of life, and relative risks are overemphasized in the medical literature.

4 Alphas and Betas



The above figure gives an example of the null hypothesis H0 and the alternative hypothesis H1 of a controlled clinical trial, in the form of t-distributions with n = 20 (H1) and its null hypothesis of no effect (H0).

According to the central limit theorem 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 ($\alpha = 0.05$ 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%. Beta (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.

5 The Main Purpose of Hypothesis Testing

The null hypothesis means, that your new treatment does not work. The main purpose of null hypothesis testing is rejecting the null hypothesis of no effect. You can reject that your new treatment does not work, but you are uncertain if it really works. Maybe it does but your sample size is too small to be sure. The alternative hypothesis means that your new treatment works. The main purpose of testing the alternative hypothesis is rejecting the alternative hypothesis. The alternative hypothesis may mean that your new treatment really works or, with adverse effects, that your adverse effect is real. With safety data analysis we try and reject the alternative hypothesis, i.e., your adverse effect is not real and can be rejected.

As adverse effects are generally yes/no data, otherwise called binary data, we must replace the t-distribution for testing with a normal distribution, which is only slightly different from the t-distribution. The 5% level is not 2.101 (dependent on sample size) but 1.96 which, in practice, is often rounded off at 2.0.

6 Limitations of Statistical Testing in General

Statistics helps you better understand the limitations of research. An important limitation is of course the presence of type I and II errors. Another limitation is the fact that statistics gives no certainty, only chances. And even less than that: it gives only *conditional* chances. Statistics only predicts chances on the understanding that. . .

For efficacy data analysis the understandings are:

the data follow normal distributions
the data are representative of your target population
your data follow the same normal distribution as that of your data
your H0 is untrue
your H1 is true.

For safety data analysis the understandings are slightly different:

the data follow normal distributions
the data are representative of your target population
your data follow the same normal distribution as that of your data
your H0 is true
your H1 is untrue.

7 FDA Rule and Guidance Classification of Adverse Effects 2012

The FDA has developed regulations based on the law and set forth in the FD&C Act (food drug and cosmetic act), a federal law enacted by Congress. It classifies two types of adverse effects. First, those that are readily interpretable as single or small numbers of events like agranulocytosis and Stevens-Johnson syndrome. They do not occur in the control group, and need no statistical test. They are called type A adverse effects. Second those, that would be anticipated to occur like for example stroke or heart attack in seniors. And, linking them to drug treatment is not self-evident, because they may well occur in untreated control groups, and a causal relationship with the drug treatment would require some statistical work up. The second class is subsequently split into two subclasses, the categories B and C adverse effects.

Categories B are adverse effects that are uncommon with drug exposure and uncommon in untreated populations. This means that a few occurrences are already suspect of causation. Nonetheless, more than one occurrence is required before causation is judged to be possible. Categories C are adverse effects that require a so-called aggregate data analysis from a clinical drug development program. These adverse effects occur more frequently in a drug treatment group than in a control group. We should add that, in order to simplify tables of adverse effect reports, the FDA lists only those serious adverse effects (SAEs), that occur with a frequency of more than 1% (or even 5% or 10%).

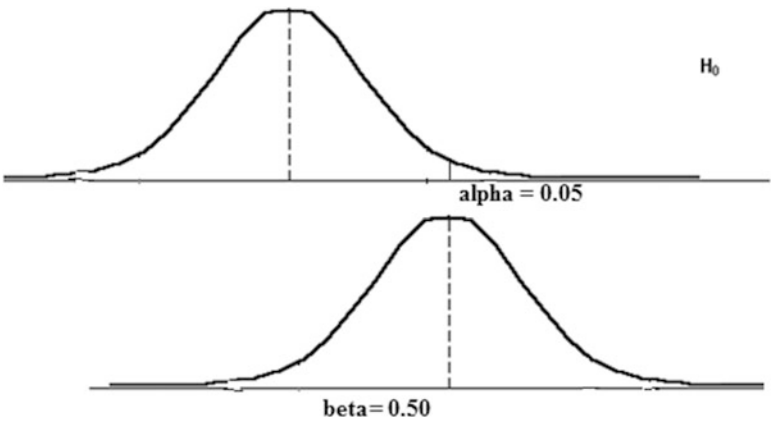
8 Emphasis on Type I Errors Is less Important with Safety Analysis

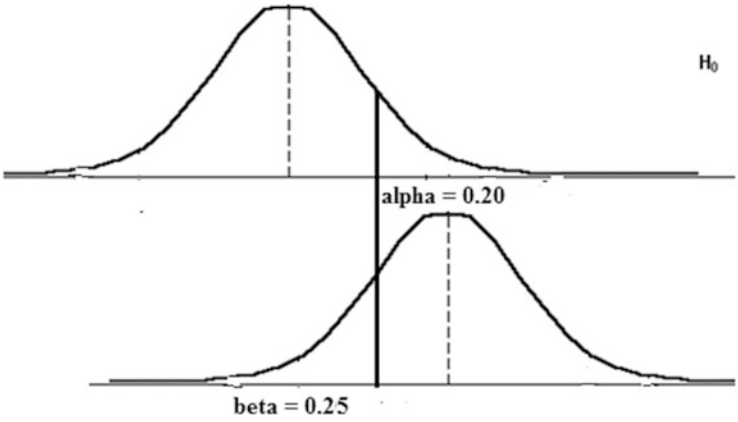
In clinical trials commonly the difference of numbers of effects in treatment and control should be small at a level of at least $p = 0.05$ or less. But with safety data, sometimes, different rules may be followed. For example, some SAEs may be reportable even if the p-value comparing the treated group to control is as high as 0.10 or even 0.20. For example, three occurrences of an SAE in a treated group and none in its equally sized control group gives a one sided p-value of 0.125. If this event is very rare and serious, and the chance of causality is clinically not negligible, an expedited report may be adequate. In contrast, other adverse effects, even if their p-values are very small, may not be reported, because the relationship with the drug is judged to be improbable and the p-value is judged to be false positive due to type I errors following from multiple testing or multiple treatments. Here, clinical judgment remains the basis of interpretation. Safety analyses, particularly those

involving summaries of common adverse effects likewise involve many and large type I errors of finding a difference where there is none, otherwise called large numbers of false positive test results. Frequentist approach of marking significant adverse effects from unadjusted p-values and unadjusted confidence intervals, although it leads to high rates of false positive results, may be better alternatives to traditional multiplicity adjustments, because they inversely tend to lead to high rates of false negative results, which are of particular concern when evaluating safety. In order to minimize the chances of both false positive and false negative results, simultaneous frequentist and Bayesian analyses are recommended for evaluating accumulating safety data. Chap. 2 gives both methods of analysis. We should add that Bayesian approaches are particularly well-suited to continuous data monitoring, because Bayesian decisions are not made on the basis of p-values but rather by summarizing posterior and prior likelihood distributions of events. In addition, Bayesian methodology allows for modeling relationships among adverse effects, removing the need for multiplicity assessment.

Emphasis on Type I error is less important for safety than for efficacy analysis. The main purpose of efficacy analysis is the rejection of the null hypothesis. You are actually trying to prove that your treatment is efficacious, and that your null hypothesis is untrue. Making the "1-alpha" area under the curve(if the null hypothesis as big as possible), is helpful to that aim. In contrast, the main purpose of the safety analysis is the rejection of the alternative hypothesis. Making the "1 - beta" area under the curve of the alternative hypothesis as big as possible is similarly helpful to the latter aim.

Maximizing the type I error may benefit safety analyses. An example of the effect of a maximized type I error on safety analysis is given underneath.





With efficacy analysis you start your statistical analysis with the prior belief, that you can reject your null hypothesis of no effect. A small type I error is a small chance of finding an effect where there is none (the type I error). With safety analysis things are different. You start with the belief of finding no adverse effect where there is one (a type II error, beta). A small type II error of finding no adverse effect where there is one, is most important here. It means a better chance of finding an adverse effect where there is one. And this is rightly so. The effects of increasing your type I error (alpha) and the effect of it on your type II error is obvious. The larger the effect of your efficacy analysis, the better the result of your study is. In contrast, the main aim of trials is not finding adverse effects. Actually, the smaller the adverse effect, the better the result of your study is.

9 Working with Flexible Alphas and Betas for Safety Analyses

Flexible alphas and betas have been recognized to be useful in efficacy analyses. For example with no life-threatening illness and a toxic compound we might wish to choose a small alpha for efficacy analysis. It means few false positives. The vertical cut-off between the area under the curve of "1-alpha" and "alpha" moves to the right. This is rightly so, because we do not want to treat the healthy. On the other hand with life threatening and no alternative we might wish to choose a small beta. It means few false negatives. The vertical cut-off between "1- alpha" and "alpha" moves to the left. We wish to treat all of the sick. With safety analyses things are slightly different. An example is given underneath.

With efficacy analyses the null hypothesis is usually rejected with a type I error of 5% and a type II error may very well be as large as 50%. With safety analysis we are more interested in smaller type II errors, because we will have increasing chance of rightly rejecting the alternative hypothesis, and that is good. After all, the main

purpose of safety analyses in clinical trials is to find no adverse effects rather than the opposite. With a type I error of 20% instead of 5% we can reject the type II error at 25%. This means that the chance of rightly rejecting the alternative hypothesis is 75% rather than 50%. This result is in better agreement with the incentive to rightly reject the alternative hypothesis, which is the main incentive of safety assessments in clinical trials.

10 Computing Minimized Betas

With larger treatment effects the type II error will be a lot smaller than 0.25 = 25%.

The value of beta in a traditional analysis of normal data can be found with the equation

$$\text{beta} = \text{prob}\ (z < \text{t-t}')$$

t = the t of the data
t′ = t yielding an area under the curve of alpha
z = an interval on the z-axis
prob. (= probability) = the area under the curve between t and t′.

With z = 2, and t′ = 5%, beta = prob of finding z < (t-t′) =
prob of finding z < (2-2) =
prob of finding z < (0) =
0.5 = 50 %.

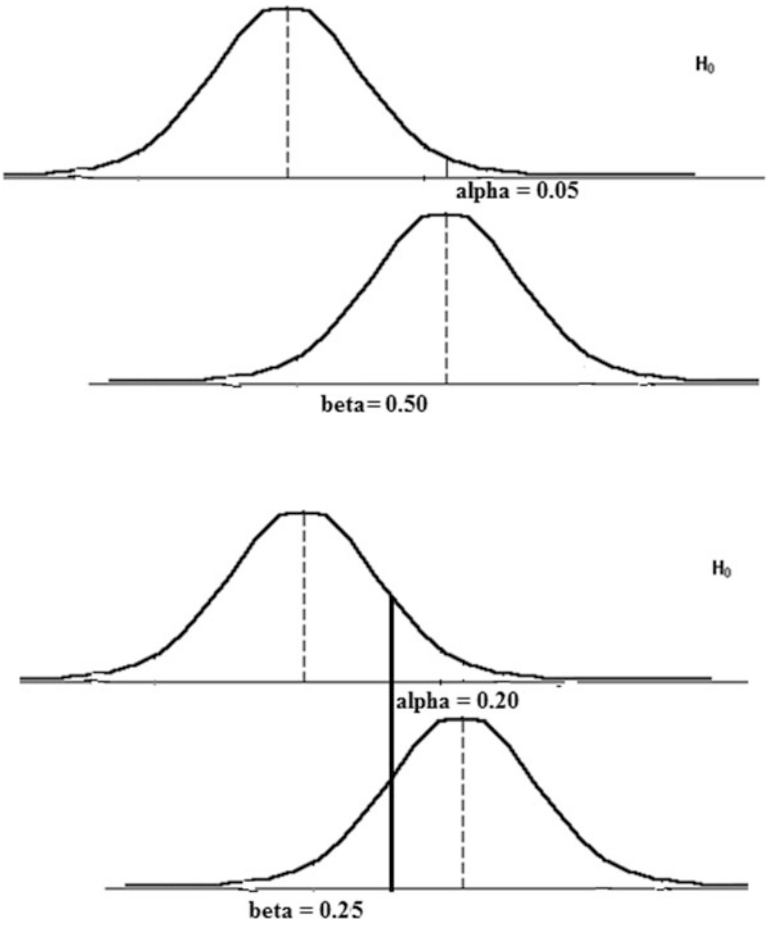
With z = 2, and t′ = 20%, beta = prob of finding z < (t-t′) =
prob of finding z < (2-0.84) =
prob of finding z < (1.16) =
0.25 = 25%

With z = 3, and t′ = 20%, beta = prob of finding z < (t-t′) =
prob of finding z < (3-0.84) =
prob of finding z < (2.16) =
0.033 = 3.3%.

With z = 4, and t′ = 20%, beta = prob of finding z < (t-t′) =
prob of finding z < (4-0.84) =
prob of finding z < (3.16) =
0.0021 = 0.21 %.

The computations can be approximated from the t-table, or from z-from-p-value calculators and from p-from-z-value calculators. A beta = type II error of only 0.21% is very small, and it means a 99.8% chance of finding an adverse effect where there is one.

11 The Effect of Increasing the Type I Error on the Magnitude of the Type II Error



The effect of increasing a type I error of 0.05 to one of 0.20 on the magnitudes of the type II error, beta, is explained more precisely in this section. The upper graph shows the null hypothesis with a vertical cut-off for a type I error (alpha) of 0.05. Underneath the corresponding alternative hypothesis is given with a type II error (beta) of 0.50. The third and fourth graphs show what will happen, if the type I error is maximized to 0.20. Now the type II error reduces from 0.50 to 0.25. Consequently, the magnitude of 1-beta = the power = the chance of finding a difference where

there is one increases from 50% to no less than 75%. This is good for safety analysis, because it means a much better chance of finding an adverse effect where there is one. However, a remaining problem is the increase of type I errors, because it means an increased chance of finding a difference where there is none. The FDA currently recommends to provide arguments in support of the study result, for example in the form of relative risks over 3, and lumping data, which is "similar data from similar studies".

12 Discussion

Flexible alphas and betas (type I and II errors) have been recognized to be useful in efficacy analyses. With efficacy analyses the null hypothesis is usually rejected with a type I error of 5% and a type II error may very well be as large as 50%. With safety analysis we are more interested in smaller type II errors, because it will give us increasing chance of rightly rejecting the alternative hypothesis, and that is good. After all, the main purpose of safety analyses in clinical trials is to find no adverse effects rather than the opposite. With a type I error of 20% instead of 5% we can reject the type II error at around 20%. This means that the chance of rightly rejecting the alternative hypothesis is around 80% rather than 50%. This result is in better agreement with the incentive to rightly reject the alternative hypothesis, which is the main incentive of safety assessments in clinical trials. For the purpose of safety analysis it is beneficial to increase the type I error, because the type II error is correspondingly reduced. This is good for safety analysis, because it means a better chance of finding an adverse effect where there is one.

In conclusion, why is forgetting about rejection of the null hypothesis and emphasizing the type II error so important with adverse drug effects. This is because in most trials for the benefit of patient safety very many possible adverse effects are being assessed. Often, standard methods for evaluation of drug safety detection is used for example the reference standard of the EU-ADR (Exploring and Understanding Adverse Drug Reactions) Consortium (Drug Safety 2013; 36: 13–23). Ninety-four drug event classes were listed, and ten of them were classified as top rank. Things will get much more complex, if we list symptoms they cause. The challenge to statistically test all of these symptoms in a meaningful way requires the simplest and most accurate statistics. We would recommend to use, in particular, 10% betas rather than 5% alphas for the very purpose.

13 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 5
Forest Plots of Adverse Effects

Abstract For the analysis of efficacy data in clinical drug trials, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportions of patients with adverse effects are assessed.

The prevalences of adverse effects can be estimated with odds values of patients with adverse effects. The chance of adverse effects of a treatment versus that of a control can be approximated from the ratios of their odds values.

Keywords Forest plots · Qualitative adverse effects · Odds ratios · Odds values

1 Introduction

For the analysis of efficacy data in clinical drug trials, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportional data have been traditionally used: proportions of patients with adverse effects (Chaps. 1 and 2).

In this chapter we will demonstrate that proportional data are increasingly replaced with odds. Odds are different from proportions. The numbers of patients with and without the illness under study in the groups 1 and 2 given underneath are a and b.

illness	yes	no
group 1	a	b
group 2	c	d

The proportion of patients with the illness in group 1 = a / (a + b).
The term proportion is synonymous to many other terms including.

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_5

the percentage of,
the risk of,
chance of,
fraction of

patients with a disease in a group. The ratio of the two proportions is often called the risk ratio. The odds of a disease is something else.

The odds of illness in group 1 = a / b.
The odds of illness in group 2 = c / d.

The odds ratio of illness in group 1 versus group 2 is given by the term a/b / c/d.

It is quite different from the risk ratio, but it can be shown that with increasingly small odds values as often observed in clinical research, odds and odds ratios tend to look increasingly similar to risks and risk ratios, and in the practice of research they are, indeed, often applied as surrogate for risk ratios. Why so. Odds are statistically easier to handle, because they run from zero to infinite, unlike risks that run from zero to 1.00. And statistical software based on "risks" often runs into a deadlock, whereas, with "odds", this is virtually never so. The same is true with ratios of odds. They are, therefore, used for assessing, whether the occurrence of adverse effects in a new treatment is significantly different from those in the control treatment. Many more novel approaches in adverse effect methodologies will be described in the current edition, for example the use of continuous data, where adverse effects are measured not only through numbers of events reported, but, instead, at a more sophisticated level through batteries of continuous data like laboratory tests. This will be demonstrated in the next chapters. In this chapter we will focus on odds values and odds ratios.

2 Systematic Assessment of Qualitative Adverse Effects

For the analysis of efficacy data in clinical drug trials, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportions of patients with side effects are assessed. As a hypothesized example of a parallel-group trial for alpha and beta blockers for the treatment of 310 patients with Raynaud's phenomenon common side effects was assessed with a questionnaire.

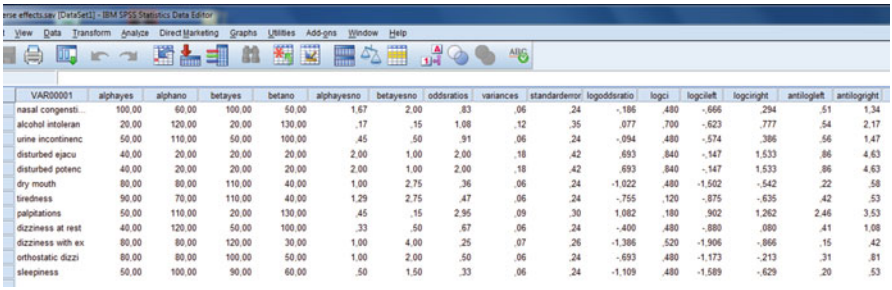
	alpha blocker n = 160		beta blocker n = 150	
side effect	yes	no	yes	no
nasal congestion	100	60	100	50
alcohol intolerance	20	120	20	130
urine incontinence	50	110	50	100
disturbed ejaculation	40	20	20	20
disturbed potence	40	20	20	20
dry mouth	80	80	110	40
tiredness	90	70	110	40
palpitations	50	110	20	130
dizziness at rest	40	120	50	100
dizziness with exercise	80	80	120	30
orthostatic dizziness	80	80	100	50
sleepiness	50	100	90	60

The above table gives numbers of patients with and without adverse effects in either of the two parallel groups. The screen view of SPSS statistics shows the results

of the Compute Variables computations of the data. The variable 00001 gives the names of the various common adverse effects in a 310 patient parallel groups studies of 160 patients on alpha blocker and 150 patients on beta blocker treated for Raynaud ‘s phenomenon. Many patients had adverse effects. An odds ratio analysis was performed to test whether the alpha blocker performed differently from the beta blocker in producing adverse effects. SPSS statistical software was applied for analysis.

3 Forest Plots of Odds Ratios

The underneath table gives the numbers of patients with and without side effect on alpha blocker and on beta blocker in the first four columns. With the help of the commands “Transform” and “Compute Variable” the odds ratios per side effect, their variances and standard errors can be computed. In order to assess whether odds ratios were significantly larger or smaller than 1.0, log transformations are required, and they can also be computed with similar commands including the log transform and antilog (exponential transforms) commands. For convenience the results of the computations are given underneath. The underneath summaries are also available as an SPSS data file in extras.springer.com, and is entitled “commonadverseeffects”.



VAR00001	alphayes	alphano	betayes	betano	alphayesno	betayesno	oddsratios	variances	standarderror	logoddsratio	logci	logcileft	logciright	antilogleft	antilogright
nasal congensti...	100,00	60,00	100,00	50,00	1,67	2,00	,83	,06	,24	-,186	,480	-,666	,294	,51	1,34
alcohol intoleran	20,00	120,00	20,00	130,00	,17	,15	1,08	,12	,35	,077	,700	-,623	,777	,54	2,17
urine incontinenc	50,00	110,00	50,00	100,00	,45	,50	,91	,06	,24	-,094	,480	-,574	,386	,56	1,47
disturbed ejacu	40,00	20,00	20,00	20,00	2,00	1,00	2,00	,18	,42	,693	,840	-,147	1,533	,86	4,63
disturbed potenc	40,00	20,00	20,00	20,00	2,00	1,00	2,00	,18	,42	,693	,840	-,147	1,533	,86	4,63
dry mouth	80,00	80,00	110,00	40,00	1,00	2,75	,36	,06	,24	-1,022	,480	-1,502	-,542	,22	,58
tiredness	90,00	70,00	110,00	40,00	1,29	2,75	,47	,06	,24	-,755	,120	-,875	-,635	,42	,53
palpitations	50,00	110,00	20,00	130,00	,45	,15	2,95	,09	,30	1,082	,180	,902	1,262	2,46	3,53
dizziness at rest	40,00	120,00	50,00	100,00	,33	,50	,67	,06	,24	-,400	,480	-,880	,080	,41	1,08
dizziness with ex	80,00	80,00	120,00	30,00	1,00	4,00	,25	,07	,26	-1,386	,520	-1,906	-,866	,15	,42
orthostatic dizzi	80,00	80,00	100,00	50,00	1,00	2,00	,50	,06	,24	-,693	,480	-1,173	-,213	,31	,81
sleepiness	50,00	100,00	90,00	60,00	,50	1,50	,33	,06	,24	-1,109	,480	-1,589	-,629	,20	,53

Variables	
1. Variable 00001	= nominal variable of different types of adverse effects
2. alphayes	= numbers of patients with adverse effects on an alpha blocker
3. alphano	= numbers of patients without adverse effects on an alpha blocker
4. betayes	= numbers of patients with adverse effect on a beta blocker
5. betano	= numbers of patients without adverse effect on a beta blocker
6. alphayesno	= odds of patients with adverse effect having alpha blocker yes/no
7. betayesno	= odds of patients with adverse effect on a beta blocker yes/no
8. oddsratios	= odds ratio of the above two odds per adverse effect
9. variances	= variance of the above odds ratio
10.standarderror	= standard error of the above adds ratio
11.logoddsratio	= logarithmetically transformed odds ratio (or)
12.logci	= logarithmetically transformed 95% confidence interval
13.logcileft	= logarithmetically transformed 95% confidence interval left end
14.logciright	= logarithmetically transformed 95% confidence interval right end
15.antilogleft	= antilog of left end term
16.antilogright	= antilog of right end term

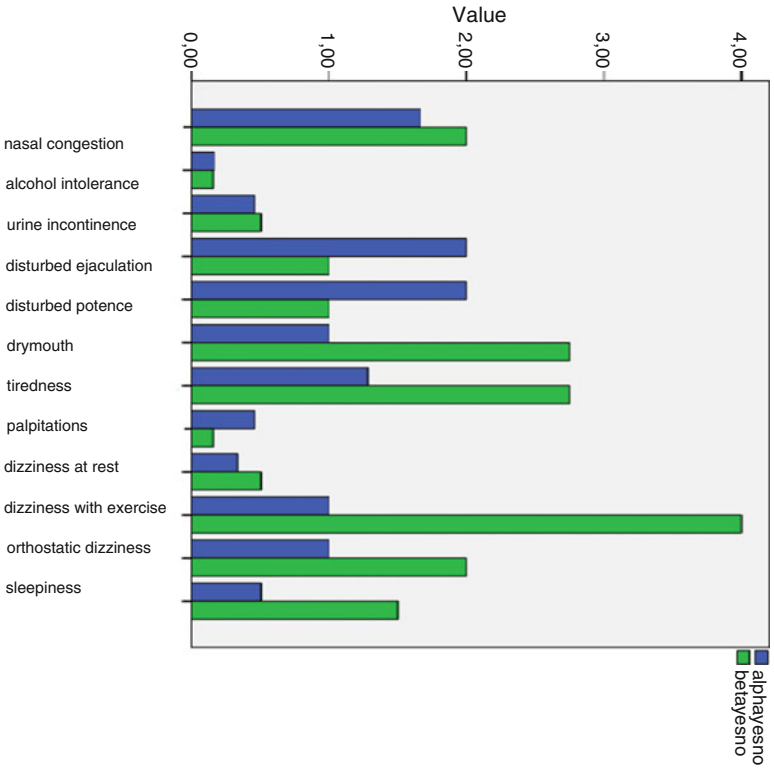
SPSS Statistical Software was used for analysis. For an overview of the odds of having had an alpha blocker, and of having had a beta blockers, the underneath commands must be given.

Command

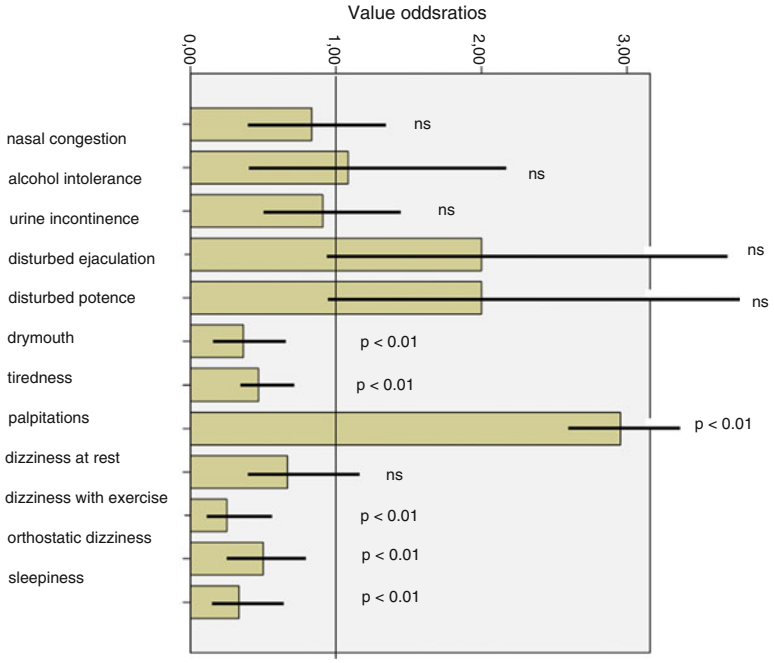
Click Graphs. . . .click Legacy Dialogs. . . .click Bars. . . .Click Clustered. . . .click Summaries of Separate Variables. . . .click variables “alphayesno” and “betayesno”. . . .enter in Bars Represent. . . .click Var00001. . . .enter in Category Axis. . . .click OK.

The underneath graph in the upright position is in the output. Copy in Word software and click in upper level of screen view the turn right command. The different side effects can be manually added with the help of Google’s Paint program.

The graph now obtained is underneath, and shows that the odds values of patients with adverse effects having alpha blockers yes/no are in blue. The odds values of parallel group patients with the same adverse effects having beta blockers are in green.



Largely the same commands can be used for the purpose of obtaining a forest plot of the odds ratios of chance of side effect on alpha blocker versus that on beta blocker. In the underneath graph again with the help of Google's Paint program different side effects as well as black fat lines of 95% confidence intervals and levels of statistical significance of the odds ratios versus an odds ration of 1.00 are drawn.



A forest plot of the odds ratios of common side effect on alpha blocker versus beta blocker is above (ns = not significantly different from an odds ratio of 1.00).

4 Discussion

For the analysis of efficacy data in clinical drug trials, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportional data have been traditionally used: proportions of patients with side effects.

In this chapter we demonstrated, that proportional data are increasingly replaced with odds. Odds are statistically easier to handle, because they run from zero to infinite, rather than zero to 1.00. The same is true with ratios of odds. They are used for assessing whether the occurrence of adverse effects in a new treatment is significantly different from those in control treatment.

A forest plot of the odds ratios of common side effect on alpha blocker versus beta blocker is above. For the analysis of efficacy data in clinical drug trials, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportions of patients with side effects are assessed.

A second novel approach in adverse effect methodology is the use of continuous data. Adverse effects may be measured through numbers of events reported, but, instead, they may equally well be estimated at a more sophisticated level with batteries of continuous data like laboratory tests. This will be demonstrated in the Chap. 6.

5 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies fifth edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and second levelers second edition, 2015,
Clinical data analysis on a pocket calculator second edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 6
Graphics of Adverse Effects

Abstract In the current chapter we will assess studies, where adverse effects are not measured through numbers of events reported, but, rather, at a more sophisticated level, with continuous data, like a battery of laboratory tests. The Konstanz information miner (Knime) was used for the graphical purpose. Include were:

box and whiskers plots,
lift charts,
histograms,
line plots,
matrices of scatter plots,
parallel coordinates,
hierarchical clusters.

Clinical computer files are complex, and hard to statistically test. Instead, visualization processes can be successfully used as an alternative approach to traditional statistical data analysis.

Keywords Konstanz information miner · Graphics of adverse effects · Box and whiskers plots · Lift charts · Histograms · Line plots · Matrices of scatter plots · Parallel coordinates · Hierarchical clusters

1 Introduction

For the analysis of efficacy data, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportions of patients with adverse effects are assessed. Instead of proportions, odds are increasingly used, because they are easier to handle as they run from zero to infinite rather than zero to 1 (unit). Similarly, the ratios of odds has computational advantages, and the availability of ratios of proportions is scanty. In the current chapter we will assess studies, where

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_6

adverse effects are not measured through numbers of events reported, but, rather, at a more sophisticated level, with continuous data, like a battery of laboratory tests.

2 Visualization Methods of Quantitative Adverse Effects

2.1 General Purpose

Computer files of clinical data are often complex and multi-dimensional, and they are, frequently, hard to statistically test. Instead, visualization processes can be successfully used as an alternative approach to traditional statistical data analysis.

For example, Knime (Konstanz information miner) software has been developed by computer scientists from Silicon Valley in collaboration with technicians from Konstanz University at the Bodensee in Switzerland, and it pays particular attention to visual data analysis. It is used since 2006 as a freely available package through the Internet. So far, it is mainly used by chemists and pharmacists, but not by clinical investigators. This section is to assess, whether visual processing of clinical data may, sometimes, perform better than traditional statistical analysis fort the detection of toxic and other adverse drug effects.

2.2 Example

Four markers for hepatotoxicity (GGT (gammagt (U/l), ASAT(asat (U/l), ALAT(alat (U/l), BILI (bilrubine (munol/l)) were measured in 150 patients. Patients were treated with an expectedly hepatotoxic compound with incrementing dosages A (low dose), B (medium dose), C (high dose). One scientific question was to assess, whether the markers could adequately predict the levels of hepatotoxicity of the incremental dosages.

Drug Dosage	asat	bili	alat	ggt	age
A	130,00	15,00	13,00	120,00	60,00
A	110,00	15,00	13,00	112,00	61,00
A	104,00	14,00	13,00	120,00	62,00
A	102,00	15,00	13,00	116,00	63,00
A	110,00	15,00	13,00	104,00	60,00
A	118,00	16,00	19,00	96,00	59,00
A	102,00	15,00	16,00	96,00	58,00
A	110,00	15,00	13,00	108,00	57,00
A	98,00	15,00	13,00	108,00	60,00
A	108,00	15,00	10,00	120,00	61,00

The data file is entitled "graphicsadverse", and is available in extras.springer.com. The first 10 patients of the 150 patient data file is above. We are interested to explore the data results for hepatotoxicity, for example, hidden data effects, like different predictive effects and frequency distributions for different subgroups. For that purpose Knime data miner will be applied. SPSS data files can not be downloaded directly in the Knime software, but excel files can, and SPSS data can be saved as an excel file (the cvs file type available in your computer must be used).

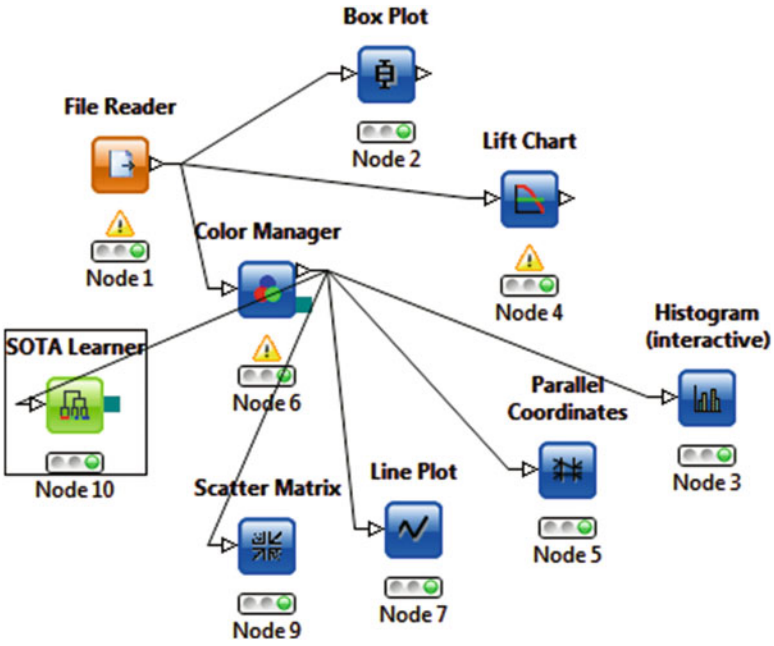
Command in SPSS
click File. . . .click Save as. . . .in "Save as" type: enter Comma Delimited (*.csv). . . . click Save.

2.3 Knime Data Miner

In Google enter the term "knime". Click Download and follow instructions. After completing the pretty easy download procedure, open the knime workbench by clicking the knime welcome screen. The center of the screen displays the workflow editor like the canvas in SPSS modeler. It is empty, and can be used to build a stream of nodes, called workflow in knime. The node repository is in the left lower angle of the screen, and the nodes can be dragged to the workflow editor simply by left-clicking. The nodes are computer tools for data analysis like visualization and statistical processes. Node description is in the right upper angle of the screen. Before the nodes can be used, they have to be connected with the "file reader" node, and with one another by arrows drawn again simply by left clicking the small triangles attached to the nodes. Right clicking on the file reader enables to configure from your computer a requested data file. . . .click Browse. . . .and download from the appropriate folder a csv type Excel file. You are set for analysis now. For convenience an CSV file entitled "graphicsadverse" has been made available at Springer's Extras Online.

2.4 Knime Workflow

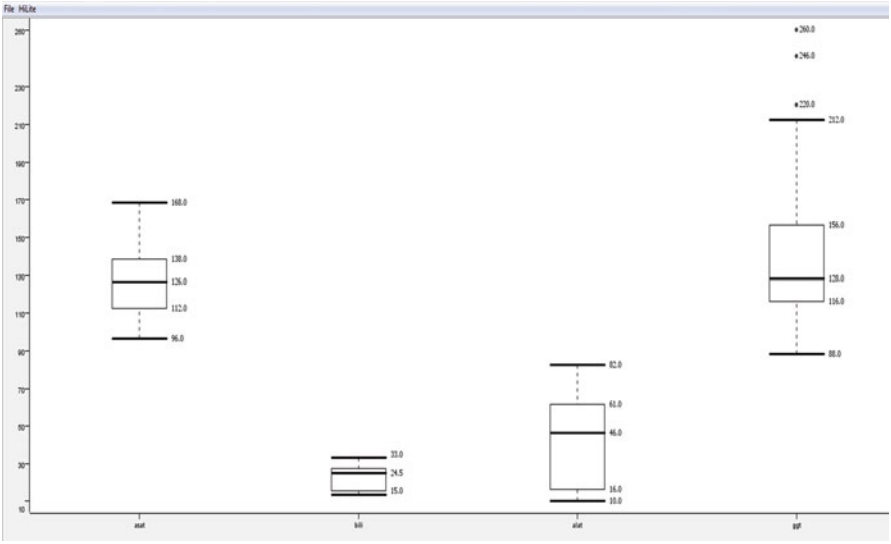
A knime workflow for the analysis of the above data example will be built, and the final result is shown in the underneath figure.



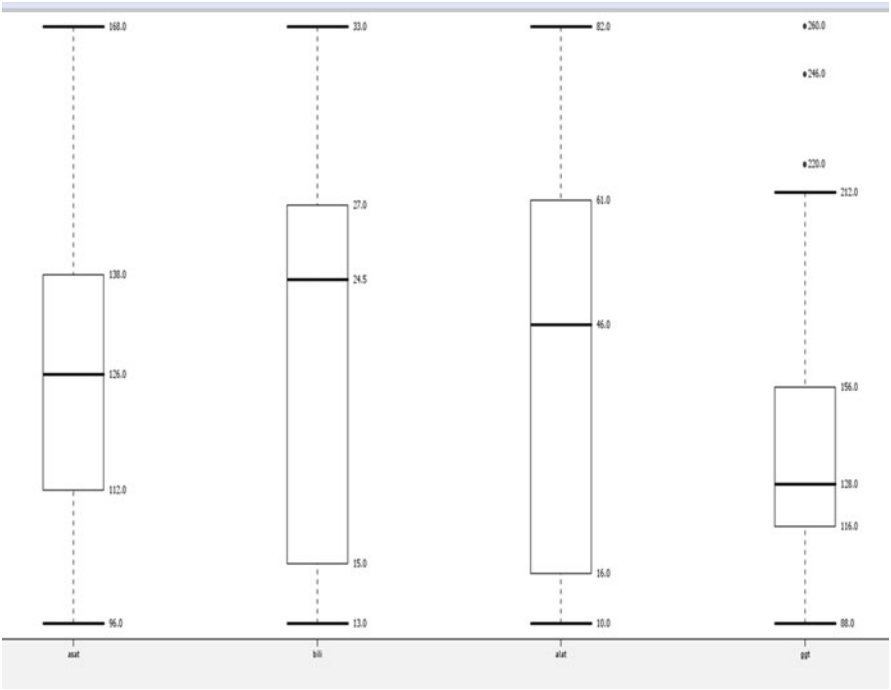
2.5 *Box and Whiskers Plots*

In the node repository find the node Box Plot. First click the IO option (import/export option nodes). Then click "Read", then the File Reader node is displayed, and can be dragged by left clicking to the workflow editor. Enter the requested data file as described above. A Node dialog is displayed underneath the node entitled Node 1. Its light is orange at this stage, and should turn green before it can be applied. If you right click the node's center, and then left click File Table a preview of the data is supplied.

Now, in the search box of the node repository find and click Data Views. . . .then "Box plot". . . .drag to workflow editor. . . .connect with arrow to File reader. . . .right click File reader. . . .right click execute. . . .right click Box Plot node. . . .right click Configurate. . . .right click Execute and open view. . . .



The above graph gives the Box plots with medians in the middle and 25% interquartile rates on either side, and with 95% confidence intervals of the four variables indicated by the whiskers on either side. The GGT plot shows that also outliers have been displayed. The smallest confidence interval has the BILI plot, and may, thus, be the best predictor of hepatotoxicity.

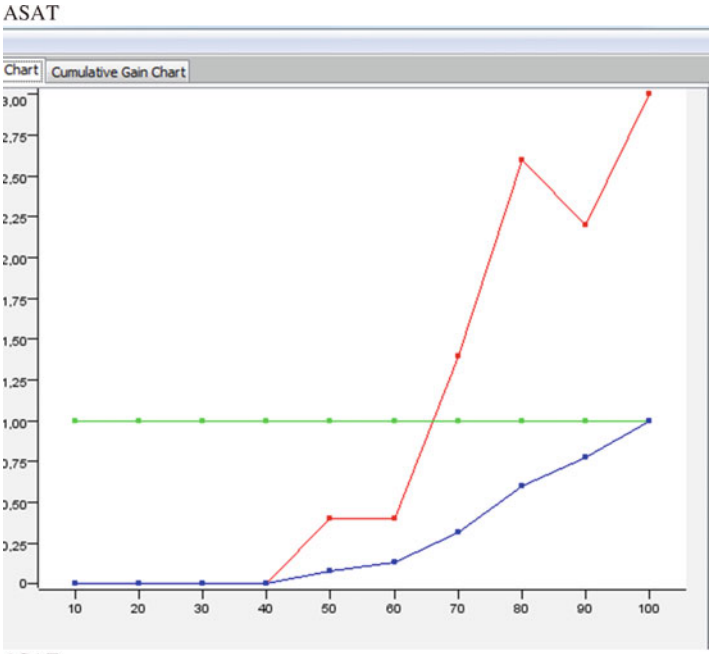


Knime also provides normalized boxplots. For the purpose minimal and maximal values are drawn in a single scale. It somewhat better visualizes the distribution pattern of each variable.

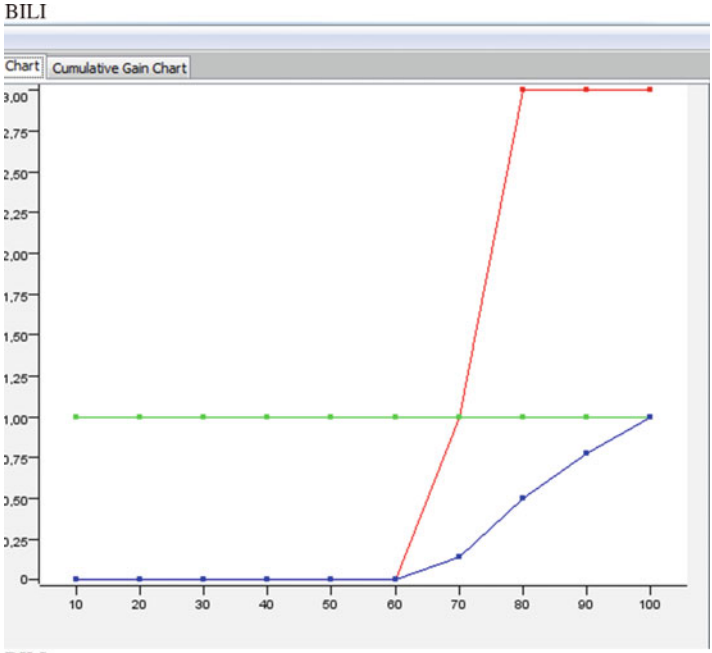
2.6 *Lift Charts*

Lift charts give a measure of effectiveness of a predictive model computed as the ratio between results obtained with and without the predictive model. In the node repository. . . .click Lift Chart and drag to workflow editor. . . . connect with arrow to File reader. . . .right click execute Lift Chart node. . . .right click Configurate. . . .right click Execute and open view. . . .

The underneath lift charts show the predictive performance of the data assuming that the four markers of liver function are predictors and the drug dosages are the outcome. If the predictive performance is no better than random, the ratio successful prediction with/without the model = 1.000 (the green line) The x-axis give dociles (1 = 10 = 10% of the entire sample etc). It can be observed underneath that at 6 or more dociles the predictive performance start to be pretty good (with ratios of 2.100–2.400). Logistic regression (here multinomial logistic regression) is being used by Knime for making predictions.

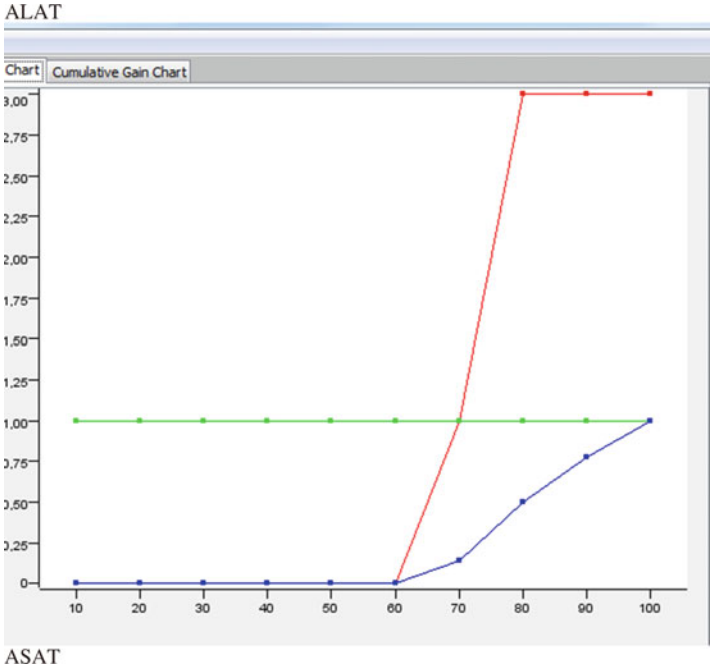


BILI



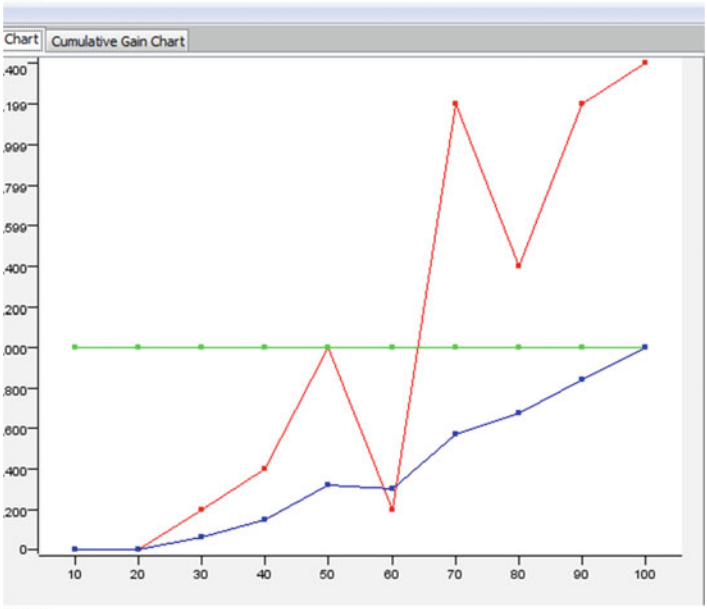
BILI

ALAT



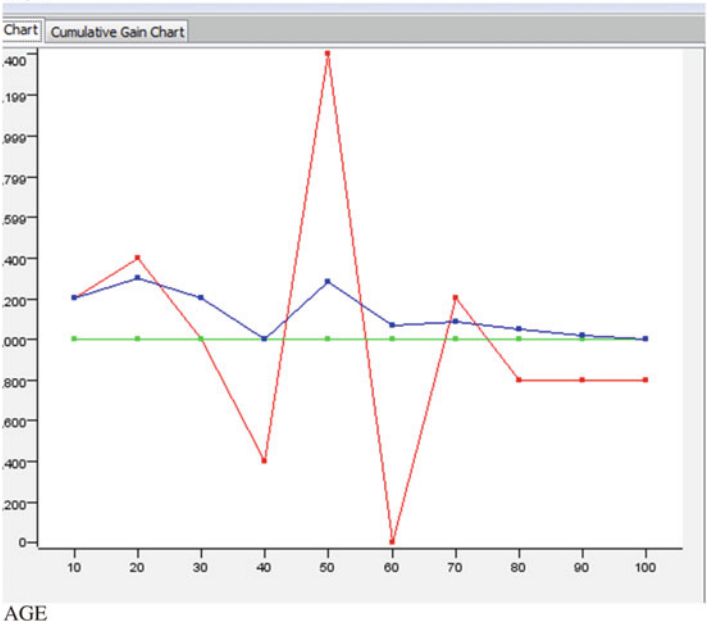
ASAT

GGT



GGT

AGE

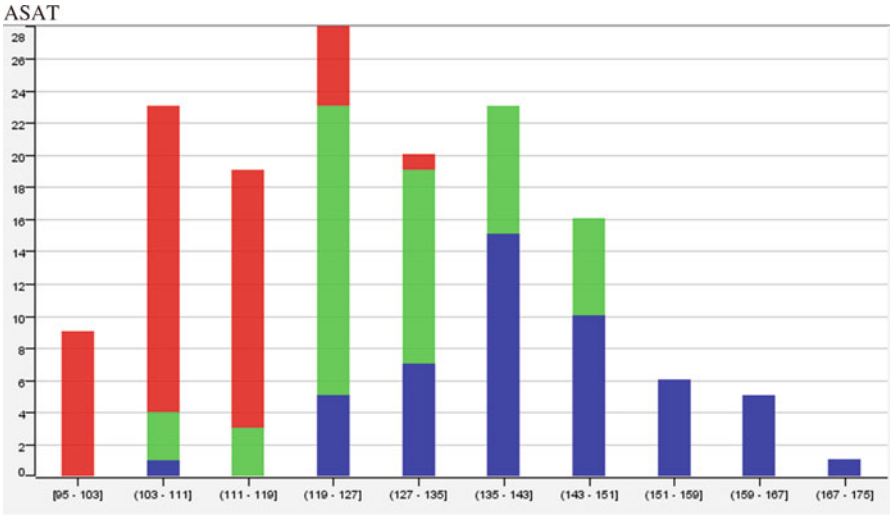


AGE

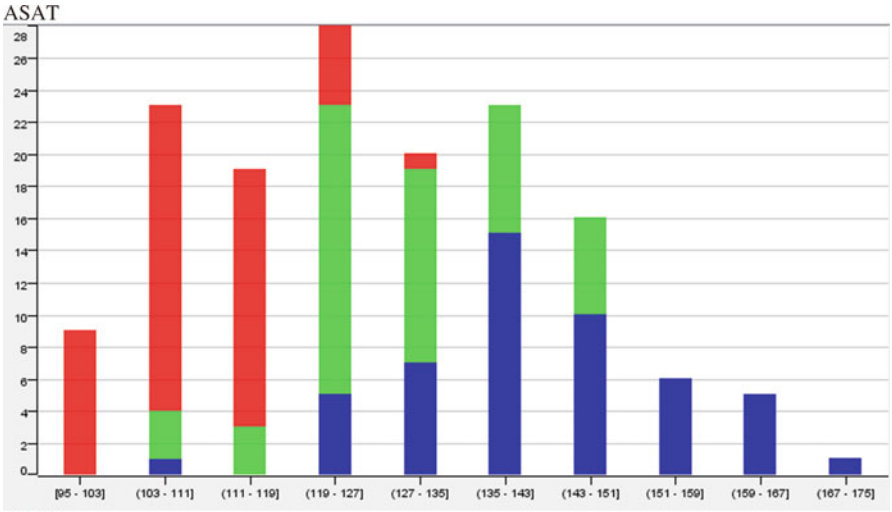
All of the lift charts, except for the last one, show that a persistent lift is obtained after 50 or so percent of the data. The age, however, does not produce a consistent lift, and, so, age is not a predictor of level of liver function.

2.7 *Histograms*

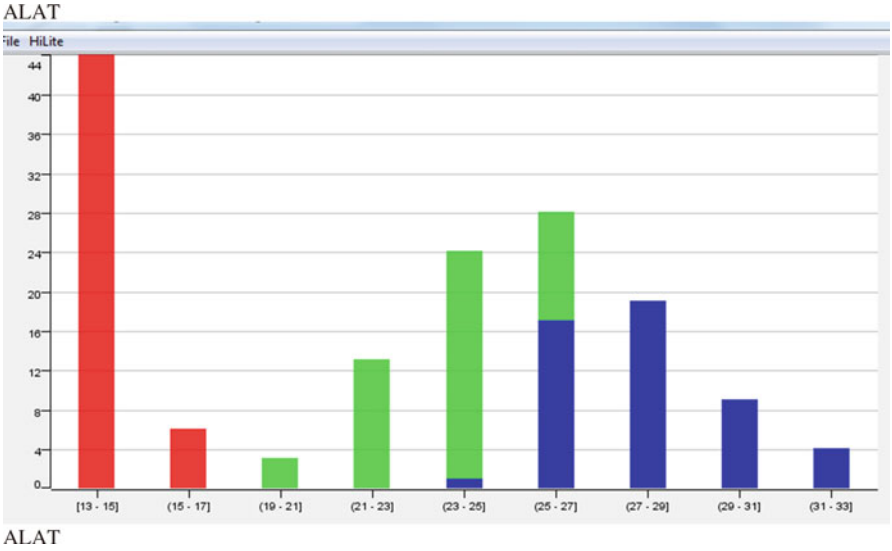
Histograms are used for accurate representation of distribution of numerical data, otherwise called probability distributions of continuous variables. In the node repository click type color. . . .click the color manager node and drag to workflow editor. . . .in node repository click color. . . .click the Esc button of your computer. . . . click Data Views. . . .select interactive histogram and transfer to workflow editor. . . . connect color manager node with File Reader. . .connect color manager with "interactive histogram node". . . .right click Configurate. . . .right click Execute and open view. . . .



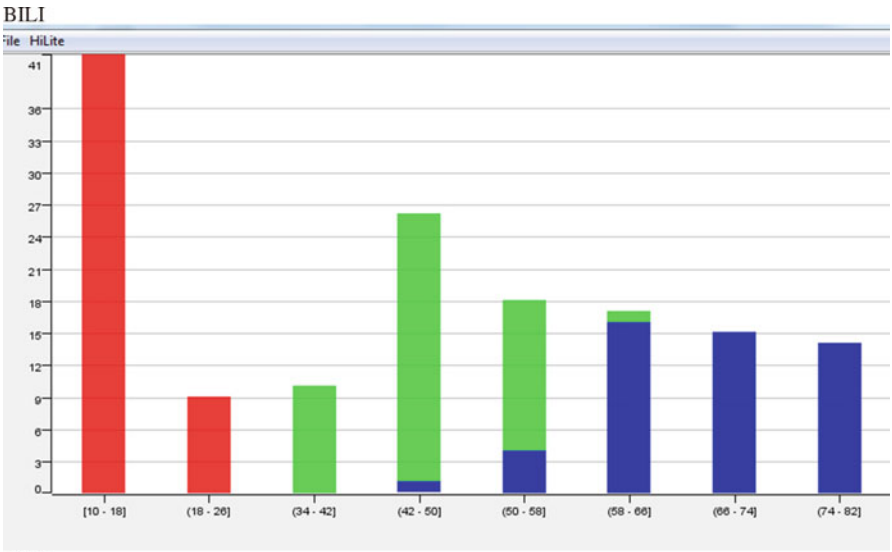
Interactive histograms with bins of ASAT values are given. The colors provide the proportions of patients with low dose (A, red), medium dose (B, green), and high dose (C, blue). It can be observed that low dose cases (red) are in ASAT 95–127 U/l cut-off. Above ASATs over 119 U/l blue (severe hepatotoxicity) is increasingly present.



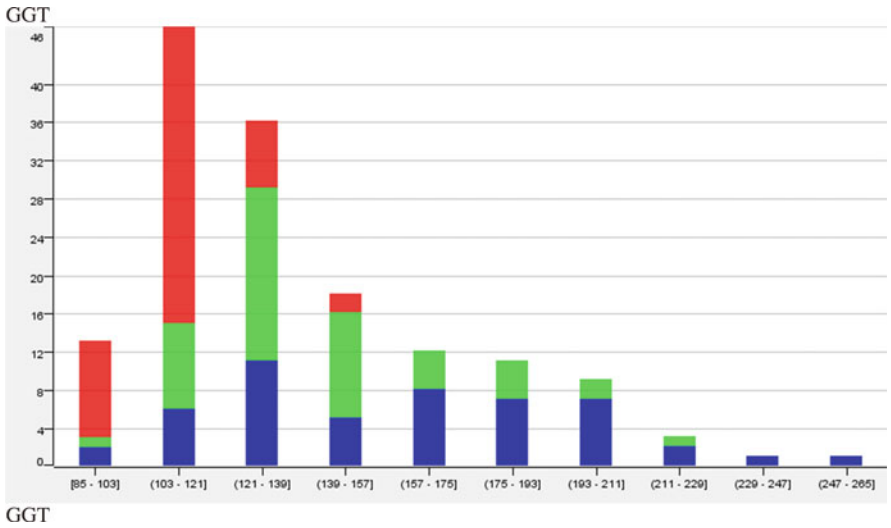
Histogram of treatments A, B, and C (red green blue). Many treatments A (red) are in the asat = 95–119 interval, many treatments B (green) are in the asat = 119–193 interval. Many treatments C (blue) are in the asat = 119–193 interval.



Histogram of treatments A, B, and C, and alat scores.



Histograms of treatments A, B, and C, and Bilirubine scores.



Histograms of treatments A, B, and C, and gammagt scores.

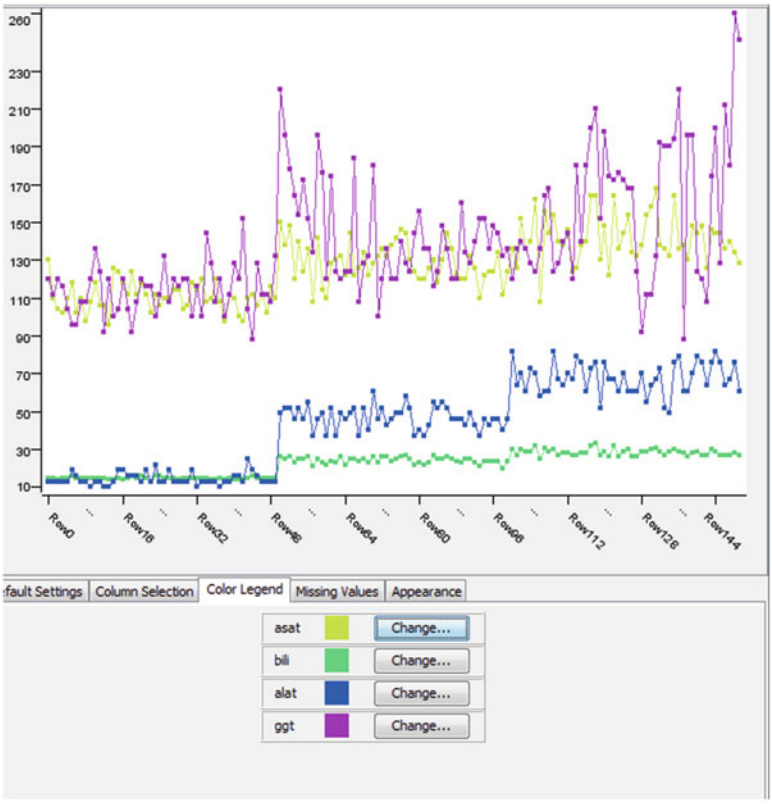
The above histograms demonstrate, how the three treatment modalities A, B, and C affect the various markers of liver function differently. Obviously, the treatment A

(red) is not very toxic and causes low levels of all 4 markers. In contrast, treatment C is more worrisome, because it induced the most increased levels of all four markers. Treatment B (green) performed in between for again all of the four markers.

2.8 *Line Plots*

Line plots can be defined as graphs of frequencies of variable values along stepped lines (called number lines). In the node repository click Data Views. . . .select the node Line plots and transfer to workflow editor. . . .connect color manager with "Line plots". . . .right click Configurate. . . .right click Execute and open view. . . .

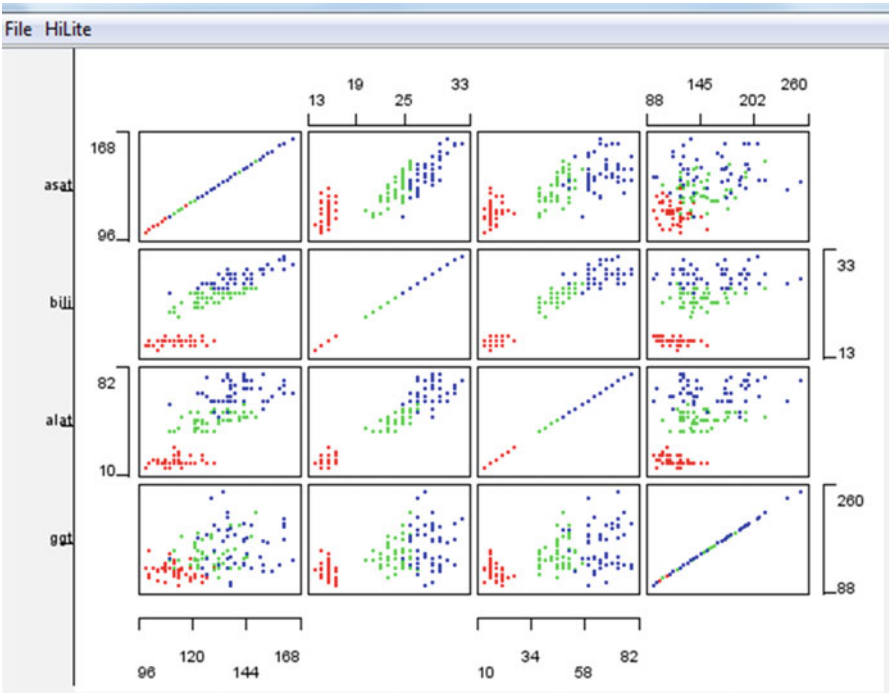
The line plot gives the values of all cases along the x-axis. The upper curve are the GGT values, The middle one the ASAT values. The lower part are ASAT and BILI values. The rows 0–50 are the cases with low drug dose, the rows 51–100 the medium dose cases, and the rows 101–150 the high dosage cases. It can be observed that particularly the GGT, ASAT, and ALAT levels increase with increasing drug dosages. This is not observed with the BILI levels.



2.9 Matrices of Scatter Plots

Matrices of scatter plots are a way to determine if you have a linear correlation between multiple variables. In the node repository click Data Views. . . .select "Matrix of scatter plots" and transfer to workflow editor. . . .connect color manager with "Matrix of scatter plots" . . . .right click Configurate. . . .right click Execute and open view. . . .

The underneath graph gives the results. The four predictors variables are plotted against one another. By the colors (blue for high dose, red for low dose) the fields show that the high dose outcomes are predominantly in the right upper quadrant, the low dose outcomes in the left lower quadrant.

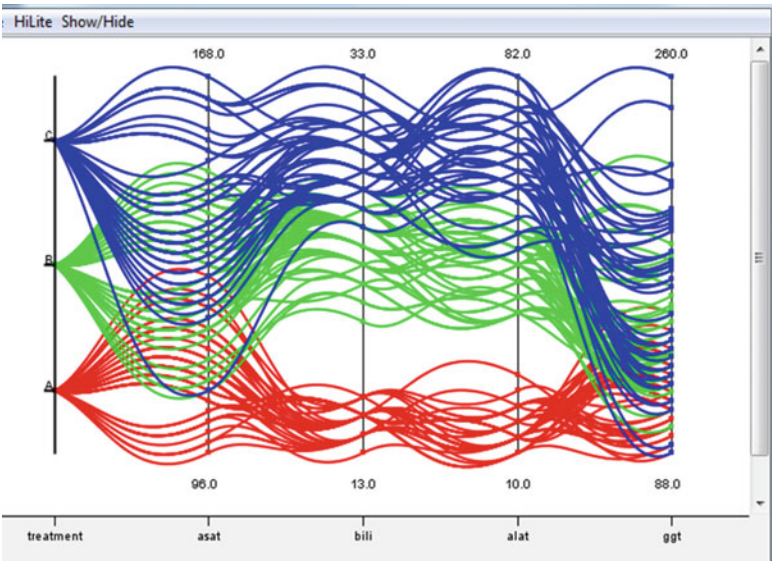


2.10 Parallel Coordinates

Parallel coordinates are a common way of visualizing high dimensional geometry and analyzing multivariate data. We will use splines instead of straight lines for even better visualization.

In the node repository click Data Views. . . .select "Parallel coordinates" and transfer to workflow editor. . . .connect color manager with "Parallel coordinates" . . . .right click Configurate. . . .right click Execute and open view. . . .click Appearance. . . .click Draw (spline) Curves instead of lines. . . .

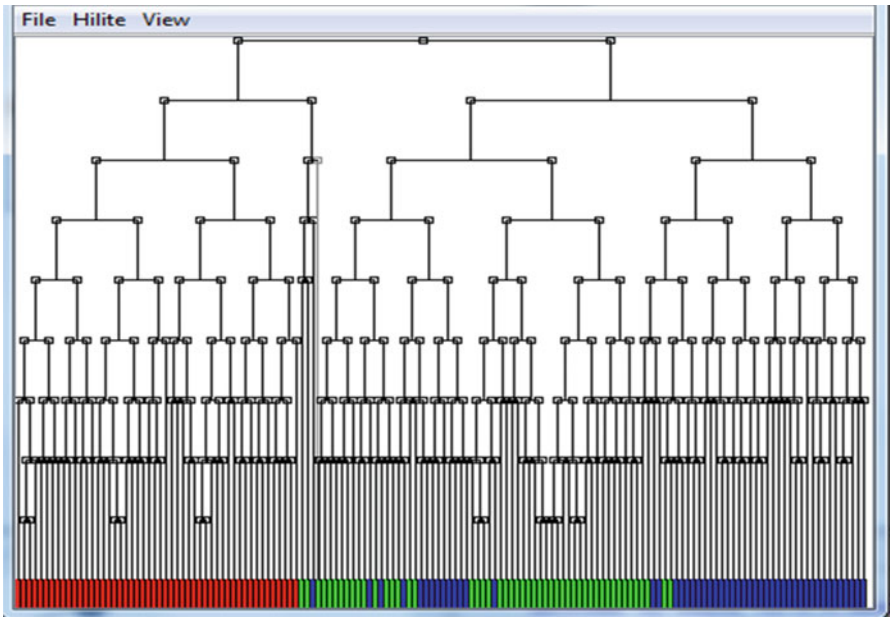
The underneath figure is given. It shows that the bili count and alat level are excellent predictors of treatment modalities. Asat are ggt also pretty good predictors of treatments A and C, however, poor predictors of levels of treatment B.



2.11 *Hierarchical Cluster Analysis*

Hierarchical cluster analysis seeks a hierarchy of distances of cluster combinations.

In the node repository click Mining. . . .select the node SOTA (Self Organizing tree Algorithm) Learner and transfer to workflow editor. . . .connect color manager with "SOTA learner". . . .right click Configurate. . . .right click Execute and open view. . . .



SOTA learning is a modified hierarchical cluster analysis, and it uses in this example the between-case distances of alat as variable. On the y-axis the standardized distances of the cluster combinations. Clicking the small squares interactively demonstrates the row numbers of the individual cases. It can be observed at the bottom of the figure that the different dosages very well cluster, with low dose (red) left, medium dose (green) in the middle, and high dose (blue) right.

3 Discussion

Clinical computer files are complex, and hard to statistically test. Instead, visualization processes can be successfully used as an alternative approach to traditional statistical data analysis. For example, Knime (Konstanz information miner) a data mining software package, mainly used by chemists and pharmacists, is able to visualize multidimensional clinical data, and this approach may, sometimes, perform better than traditional statistical testing. In the current example it was able to demonstrate the clustering of hepatic markers to identify the effect of incremental hepatotoxic drug dosages. More background, theoretical and mathematical information of splines and hierarchical cluster modeling are in Machine learning in medicine part one, Chap. 11, Non-linear modeling, pp 127–143, and Chap. 15, Hierarchical cluster analysis for unsupervised data, pp 183–195, Springer Heidelberg Germany, from the same authors.

For the analysis of efficacy data in clinical drug trials, generally, continuous data are applied. For the analysis of safety data, in contrast, usually proportions of patients with adverse effects are assessed.

We already reviewed in the Chaps. 2 and 3 traditional assessments of common adverse effect with tables of proportional data. Instead of proportions odds are increasingly used. They are statistically easier to handle, because they run from zero to infinite, rather than zero to 1.00. Similarly the ratios of odds has computational advantages, the software of ratios of proportions are scanty, and may easily run aground. In the current chapter we assessed studies where adverse effects were not measured through numbers of events reported, but rather at a more sophisticated level with continuous data like a battery of laboratory tests. In the next chapter we will assess adverse effects in more longitudinal studies adjusted for time effects, rather than those with single endpoint measurements. Mixed linear models for better sensitivity of testing will be used for the purpose.

4 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 7
Adverse Effects in Clinical Trials with Repeated Measures

Abstract This chapter assesses adverse-effect-analyses in more longitudinal studies adjusted for time effects rather than those with single endpoint measurements. Mixed linear models are particularly suitable for the purpose, because within-subject differences receive fewer degrees of freedom, than they do with traditional general linear models. The mixed effect model in our example was able to demonstrate a significant adverse effect of the independent type, while the classical repeated measures analysis was unable to do so. In the data example the adverse effect was, thus, significantly more present in one group than it was in the other.

Keywords Longitudinal studies · Mixed linear models · Repeated measures analysis of variance

1 Introduction

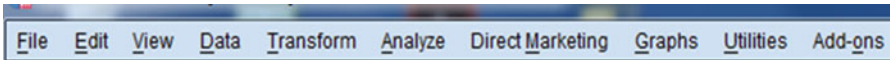
We will assess adverse-effect-analyses in more longitudinal studies adjusted for time effects rather than those with studies of single endpoint measurements. The former studies include repeatedly measured outcomes, and should be adjusted for time dependent within-subject differences in adverse effects. The adjustment is, because of differing within-subject correlation levels with repeated measurements. Mixed linear models will be particularly suitable for the purpose, because within-subject differences receive fewer degrees of freedom, than they do with traditional general linear models, since they are nested into a separate layer or subspace. In this way better sensitivity is left in the model to demonstrate differences between-subjects. In trial models, where the occurrence of adverse effects in treatment and control are compared, and where the outcome has multiple measurements in one subject, the aim of your research is to demonstrate differences between-subject, rather than within-subject, and a mixed model is, thus, a better choice.

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_7

2 Data Example, Mixed Linear Models

In a parallel-group study of 20 patients with glaucoma the effect on eye pressure of a novel beta-blocker as compared to a standard beta-blocker was assessed. Also fall in heart rate (beats per minute) was measured, because severe bradycardia was the main adverse effect expected. This adverse effect was assumed to be independent of the main outcome of the study, eye pressure, but we wished to know, whether this adverse effect was or was not significantly different-sized between the treatment modalities.

The fall in heart rate was measured each week. Underneath the data screen of the measured bradycardias is given in SPSS statistical software. The entire data file is also given in Springer Extras Online and is entitled "sideeffectmixed".

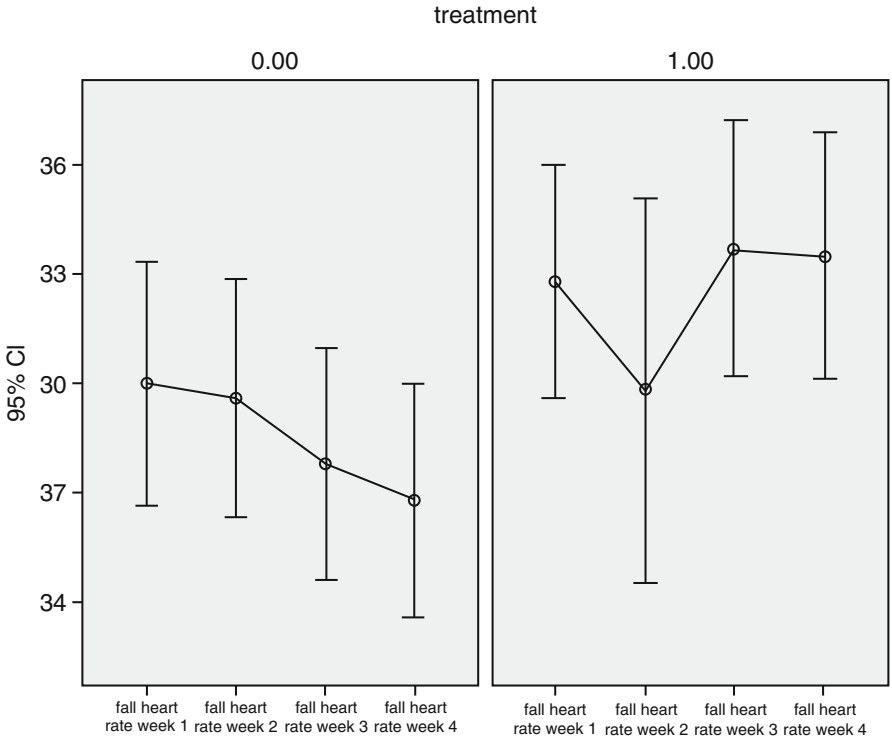


7 :

	id	treatment	firstweek	secondweek	thirdweek	fourthweek
1	1	,00	22,00	23,00	21,00	20,00
2	2	,00	24,00	23,00	22,00	21,00
3	3	,00	28,00	28,00	26,00	25,00
4	4	,00	30,00	29,00	27,00	26,00
5	5	,00	30,00	29,00	27,00	26,00
6	6	,00	31,00	30,00	28,00	27,00
7	7	,00	31,00	30,00	29,00	28,00
8	8	,00	31,00	31,00	29,00	28,00
9	9	,00	36,00	36,00	34,00	33,00
10	10	,00	37,00	37,00	35,00	34,00
11	11	1,00	26,00	21,00	26,00	26,00
12	12	1,00	27,00	22,00	27,00	27,00
13	13	1,00	31,00	26,00	32,00	32,00
14	14	1,00	32,00	27,00	33,00	33,00
15	15	1,00	32,00	46,00	33,00	33,00
16	16	1,00	33,00	28,00	34,00	34,00
17	17	1,00	34,00	29,00	35,00	34,00
18	18	1,00	34,00	29,00	35,00	35,00
19	19	1,00	39,00	34,00	41,00	40,00
20	20	1,00	40,00	36,00	41,00	41,00

Start by opening the data file in your computer mounted with SPSS statistical software. First, commands for a graph of the levels of bradycardias will be given.

Command
click Graphs. . . .click Legacy Dialogs. . . .click Error Bars. . . .click Simple. . . .mark Summaries of Separate Variables. . . .click Define. . . .Error Bars: enter bradycardia, week 1, week 2, week 3, week 4. . . .Bars Represent: choose Confidence Interval for mean 95%. . . .click Columns: enter Treatment. . . .click OK.



The above graphs are in the output sheets. Mean fall in heart rate and 95% confidence intervals after treatment 0 and treatment 1 have been measured. In order to test whether the differences between the two treatments are significant, next, a repeated measures analysis of variance (anova) is performed.

Command
click General Linear Model. . . .Repeated Measures. . . .Within-Subject Factor Number. . . .enter week: Number of Levels: 4. . . .click Add. . . .Measure Name: repeated. . . .click Define. . . .Within-Subject Variables: enter week 1, 2, 3, 4. . . . Between-Subject Factor(s): enter Treatment. . . .click OK.

The underneath tables are in the output sheets.

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
week	Pillai's Trace	,900	47,745[a]	3,000	16,000	,000
	Wilks' Lambda	,100	47,745[a]	3,000	16,000	,000
	Hotelling's Trace	8,952	47,745[a]	3,000	16,000	,000
	Roy's Largest Root	8,952	47,745[a]	3,000	16,000	,000
week * treatment	Pillai's Trace	,941	84,649[a]	3,000	16,000	,000
	Wilks' Lambda	,059	84,649[a]	3,000	16,000	,000
	Hotelling's Trace	15,872	84,649[a]	3,000	16,000	,000
	Roy's Largest Root	15,872	84,649[a]	3,000	16,000	,000

a. Exact statistic
b. Design: Intercept + treatment
Within Subjects Design: week

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
week	,001	111,572	5	,000	,346	,369	,333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.
b. Design: Intercept + treatment
Within Subjects Design: week

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
week	Sphericity Assumed	32,700	3	10,900	2,411	,077
	Greenhouse-Geisser	32,700	1,038	31,501	2,411	,137
	Huynh-Feldt	32,700	1,106	29,564	2,411	,134
	Lower-bound	32,700	1,000	32,700	2,411	,138
week * treatment	Sphericity Assumed	133,700	3	44,567	9,859	,000
	Greenhouse-Geisser	133,700	1,038	128,799	9,859	,005
	Huynh-Feldt	133,700	1,106	120,879	9,859	,004
	Lower-bound	133,700	1,000	133,700	9,859	,006
Error(week)	Sphericity Assumed	244,100	54	4,520		
	Greenhouse-Geisser	244,100	18,685	13,064		
	Huynh-Feldt	244,100	19,909	12,261		
	Lower-bound	244,100	18,000	13,561		

Tests of Between-Subjects Effects

Measure:repeated
Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	74420,000	1	74420,000	851,433	,000
treatment	304,200	1	304,200	3,480	,078
Error	1573,300	18	87,406		

All of the multivariate tests are very significant. Obviously, a multivariate analysis of variance with all of the weeks of treatment as outcome is, overall, statistically significant. This means, that there is in this model a statistical significance, but we don't know exactly where. Mauchly's test is significant, which means, that the hypothesis of equal variances in the groups is not warranted. The within-subject test assuming sphericity (≈ equal variances) is, therefore, not adequate. However, the three additional within-subject tests can be adequately applied, but none of them were significant. The test of difference between the treatments is neither statistically significant. And so, we have to conclude, that the null hypothesis of no difference between the two treatment effects can not be rejected. Subsequently, a mixed linear analysis of the study will be performed. For the purpose the data file as it currently is, will have to be restructured.

Command

click Data. . . .click Restructure. . . .mark Restructure selected variables into cases. . . . click Next. . . .mark One (for example, w1, w2, and w3). . . .click Next. . . .Name: id (the patient id variable is already provided). . . .Target Variable: enter "firstweek, secondweek. . .. . . fourthweek". . . .Fixed Variable(s): enter treatment. . . .click Next. . . . How many index variables do you want to create?. . . . mark One. . . .click Next. . . .click Next again. . . .click Next again. . . .click Finish. . . . Sets from the original data will still be in use. . .click OK.

Return to the main screen, and observe that there are now 80 rows instead of 20 in the data file. The table is now adequate to perform a mixed linear model analysis. For readers' convenience it is saved in extras.springer.com, and is entitled "sideeffectmixedrestructured". SPSS calls the levels "indexes", and the outcome values after restructuring "Trans" values, terms pretty confusing to us. Click the data screen. It now looks like given underneath. Index1 is the week of treatment, trans1 is the outcome value, the level of bradycardia per patient per week.

id1	treatment	Index1	trans1
1	,00	1	22,00
1	,00	2	23,00
1	,00	3	21,00
1	,00	4	20,00
2	,00	1	24,00
2	,00	2	23,00
2	,00	3	22,00
2	,00	4	21,00
3	,00	1	28,00
3	,00	2	28,00
3	,00	3	26,00
3	,00	4	25,00
4	,00	1	30,00
4	,00	2	29,00
4	,00	3	27,00
4	,00	4	26,00
5	,00	1	30,00
5	,00	2	29,00
5	,00	3	27,00
5	,00	4	26,00
6	,00	1	31,00
6	,00	2	30,00
6	,00	3	28,00
6	,00	4	27,00
7	,00	1	31,00
7	,00	2	30,00
7	,00	3	29,00
7	,00	4	28,00
8	,00	1	31,00
8	,00	2	31,00
8	,00	3	29,00
8	,00	4	28,00
9	,00	1	36,00
9	,00	2	36,00
9	,00	3	34,00
9	,00	4	33,00
10	,00	1	37,00
10	,00	2	37,00
10	,00	3	35,00
10	,00	4	34,00
11	1,00	1	26,00
11	1,00	2	21,00
11	1,00	3	26,00

11	1,00	4	26,00
12	1,00	1	27,00
12	1,00	2	22,00
12	1,00	3	27,00
12	1,00	4	27,00
13	1,00	1	31,00
13	1,00	2	26,00
13	1,00	3	32,00
13	1,00	4	32,00
14	1,00	1	32,00
14	1,00	2	27,00
14	1,00	3	33,00
14	1,00	4	33,00
15	1,00	1	32,00
15	1,00	2	46,00
15	1,00	3	33,00
15	1,00	4	33,00
16	1,00	1	33,00
16	1,00	2	28,00
16	1,00	3	34,00
16	1,00	4	34,00
17	1,00	1	34,00
17	1,00	2	29,00
17	1,00	3	35,00
17	1,00	4	34,00
18	1,00	1	34,00
18	1,00	2	29,00
18	1,00	3	35,00
18	1,00	4	35,00
19	1,00	1	39,00
19	1,00	2	34,00
19	1,00	3	41,00
19	1,00	4	40,00
20	1,00	1	40,00
20	1,00	2	36,00
20	1,00	3	41,00
20	1,00	4	41,00

Now, a mixed linear analysis will be performed of the above table.

Command

Analyze. . . .Mixed Models. . . .Linear. . . .Specify Subjects and Repeated. . . .Subject: enter id . . . .Continue. . . .Linear Mixed Model. . . .Dependent Variables: Trans1 . . . .Factors: Index1, treatment. . . .Fixed. . . .Build Nested Term. . . .Treatment . . . .Add . . . .Index1. . . .Add. . . . Index1 build term by* treatment. . . .Index1 *treatment. . . . Add . . . .Continue. . . .click OK (* = sign of multiplication).

In the output sheets the underneath tables are given.

Information Criteria[a]

-2 Restricted Log Likelihood	452,829
Akaike's Information Criterion (AIC)	460,829
Hurvich and Tsai's Criterion (AICC)	461,426
Bozdogan's Criterion (CAIC)	473,936
Schwarz's Bayesian Criterion (BIC)	469,936

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: bradycardia.

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	72	2948,300	,000
treatment	1	72	12,052	,001
Index1	3	72	,432	,731
Index1 * treatment	3	72	1,766	,161

a. Dependent Variable: fall heart rate week 1.

The above results of the mixed linear model show, that one treatment modality performs better than the other at $p = 0.001$, and this is adjusted for within-subjects differences between repeated measures. In conclusion, the traditional repeated measures analysis of variance was not sensitive to detect a between-subject difference of two parallel treatments. After adjustment for the error due to differences within the subjects due to the repeated measures, using a nested approach, the mixed linear model enabled to observe a very significant difference between the adverse effects of the treatments 0 and 1, at $p < 0.001$.

3 Discussion

We assessed adverse-effect-analyses in more longitudinal studies, adjusted for time effects, rather than those in cross-sectional studies with single endpoint measurements. The former studies include repeatedly measured outcomes, and have been adjusted for time dependent within-subject differences in adverse effects. This adjustment is done, because of differing within-subject correlation levels with repeated measurements. Mixed linear models are suitable for the purpose, because within-subject differences receive fewer degrees of freedom, than they do with traditional general linear models, since they are nested into a separate layer or subspace. In this way better sensitivity is left in the model to demonstrate differences between-subjects. In trial models, where the occurrence of adverse effects in treatment and control are compared, and where the outcome has multiple measurements in one subject, the aim of your research is to demonstrate differences between-subject, rather than within-subject, and a mixed model is, thus, a better choice.

In conclusion, treatment adjustment for within-subject differences did not significantly change bradycardia rates in the traditional repeated measures anova. With mixed models a better sensitivity of testing was left in the model to demonstrate differences between-subjects. Therefore, if the main aim of your research is to demonstrate differences between-subjects, a mixed model was a better choice. The mixed effect model in our example was able to demonstrate a significant adverse effect of the independent type, while the classical analysis was unable to do so. In the data example the adverse effect was significantly more present in one group than it was in the other.

4 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 8
Benefit Risk Ratios

Abstract The current chapter gives an example of how, with the help of advanced statistics, pretty convincing quantitative conclusions can be drawn about drug risk and safety measurements.

For the purpose confidence intervals of ratios of normal variables are needed. Using the quadratic method, these confidence intervals can be approximated. In order to compute the variance of the ratio of two paired terms the underneath equation is required. In an example assessing the chance of intolerable dizziness of different treatment modalities for angina pectoris, the benefit risk ratio was estimated from

$$\text{Benefit/risk ratio} = \frac{\text{improved exercise tolerance}}{\text{reduced patient compliance}}$$

Benefit/risk ratios are the first step to a more quantitative approach to the analysis of drug safety.

Keywords Benefit risk ratios · Confidence intervals of ratios · Quadratic method · Paired benefit risk assessment · Quantitative analysis of drug safety

1 Introduction

Benefit risk assessment is the basis for the FDA ‘s (American food drug administration) ongoing monitoring and regulatory reviews of human drugs and biologics, and on March 30 2018 a novel benefit risk implementation plan in drug regulatory decision making was published in FDA.GOV, FY 2018–2022, entitled Benefit-risk assessment in drug regulatory decision making. Despite this document of over 7000 words emphasizing the need for enhanced and better communicated assessments and standardized operating procedures for identifying benefits outweighing risks, and

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_8

providing extensive evidence about drug's safety, the entire document makes no quantitative claims, and states, in its final lines, that the benefit risk framework the FDA works with, is no more than a structured *qualitative* approach. But it is promised, that systematic ways for more quantitative benefit risk analyses will be explored. The current chapter gives an example of how, with the help of advanced statistics, pretty convincing quantitative conclusions can be drawn about drug risk and safety measurements. For the purpose confidence intervals of ratios of normal variables are required. Using the quadratic method, these confidence intervals can be approximated. In order to compute the variance of the ratio of two paired terms the underneath equation is adequate.

$$\text{Variance}\,(x + y) = \text{variance}\,(x) + \text{variance}\,(y) + 2\,\text{covariance}\,(x, y)$$

The variance of the ratio of two paired terms can also be approached from a Taylor series using the delta method.

$$\text{Variance}\,(x + y) = y^2\,\text{variance}\,(x) + x^2\text{variance}\,(y)$$

By combining the equations we will end up finding an adequate estimate of a drug risk and safety ratio, at least, if variances are not too small, and samples are not too small. As an example a trial performed by our group, and published in the Br J Clin Pharmacol (1991; 50: 545–60) was used.

2 Example

In a 335 patient double blind controlled parallel-group trial from our group (Br J Clin Pharmacol 1991; 50: 545–60), in patients with coronary artery disease the effect of different dosages of calcium channel blockers was assessed on time to ischemia during an exercise tolerance test, as efficacy variable, and percentage of withdrawals from study due to intolerable dizziness, as safety variable. The ratio of the two was used as benefit/risk estimator of treatments with three different types of calcium channel blockers, namely amlopidine 5 and 10 mg, diltiazem 200 and 300 mg, and mibefradil 50 and 100 mg daily.

We included male and female outpatients, between 18 and 75 years of age, suffering from chronic stable angina pectoris and receiving stable β-adrenoceptor blocker treatment (heart rate at rest between 55 and 70 beats min^{-1}) for at least 2 weeks before the start of the study treatments. Patients were eligible when reproducible myocardial ischemia occurred during exercise testing (ETT). Underneath an exercise performance table is given.

	Amlodipine		*Diltiazem*		*Mibefradil*	
	5 mg	*10 mg*	*200 mg*	*300 mg*	*50 mg*	*100 mg*
Rest						
SBP (mmHg)	−6±19	−9±20	1±20	−3±10	−3±9	−6±10
DBP (mmHg)	−3±9	−5±10	−2±9	−3±10	−3±9	−6±10
HR (beats min^{-1})	1±8	3±10	−2±7	−2±10	−7±8	−13±8
RPP (mmHg beats min^{-1} 10^{-3})	−0.2±1.5	−0.2±1.9	−0.1±1.7	−0.3±1.9	−1.1±1.3	−1.9±1.6
PQ-interval (ms)	164±23	166±23	170±23	171±23	166±24	173±27
QTc (ms)	404±31	405±35	402±35	407±35	395±33	393±33
Onset of ischaemia						
SBP (mmHg)	3±25	1±20	3±22	1±26	0±23	−4±23
DBP (mmHg)	−2±9	−4±10$^{\chi}$	−2±10	−3±10‡	−1±13	−7±11
HR (beats min^{-1})	−1±12$^{\chi\chi}$	0±12†$^{\chi}$	−5±11	−7±14†	−6±15	−18±20
RPP (mmHg beats min^{-1} 10^{-3})	0.3±4.1	0.1±3.5‡	−0.6±3.5	−1.1±4.6†	−1.1±4.6	−3.4±5.1
Maximal workload						
SBP (mmHg)	3±22	0±23	2±21	1±24	−2±23	−4±19
DBP (mmHg)	−2±11	−4±11$^{\chi}$	−2±11	−3±10‡	−2±12	−7±12
HR (beats min^{-1})	0±9¶$^{\chi\chi}$	−1±13$^{\chi\chi}$¶	−5±10‡	−8±11‡	−15±14	−25±16
RPP (mmHg beats min^{-1} 10^{-3})	0.3±3.7‡	−0.2±3.9$^{\chi\chi}$	−0.7±3.7$^{\chi}$	−1.3±4.0‡	−2.8±4.3	−5.1±4.3
Time to onset of angina (s)	16±46	25±56	20±47	31±69	16±57	30±64
Time to onset of ischaemia (s)	25±50‡	33±63‡	30±64†	39±70‡	54±55	70±62

ETT = exercise testing, SBP = systolic blood pressure, DBP = diastolic blood pressure, HR = heart rate, RPP = rate pressure product, QTc = corrected QT-interval.
$^{\chi}P<0.03$ *vs* mibefradil at comparable dosage. $^{\chi\chi}P<0.003$ *vs* mibefradil at comparable dosage. †$P<0.02$ *vs* mibefradil at comparable dosage.
‡$P<0.001$ *vs* mibefradil at comparable dosage. ¶$P<0.003$ *vs* diltiazem at comparable dosage.
*Changes from baseline are presented, except for the PQ interval and QTc that are described in terms of absolute measurements.

3 Benefit/Risk Analysis

$$\text{Benefit/risk ratio} = \frac{\text{improved exercise tolerance}}{\text{reduced patient compliance}}$$

Serious symptoms of dizziness occurred in up to 14% of the randomized patients, and caused 19 patients treated with mibefradil to withdraw from the study. The underneath table left column shows the improved exercise tolerance as estimated with percentages increased time to onset of ischemia during ETTs (exercise tolererance tests), and reduced patient compliance as estimated with percentages of withdrawals due to symptoms of dizziness, and their ratios. In the second column from the left are the same effects after 10 days on amlodipine 10 mg, etc. The ratios of improved exercise tolerance and reduced patient compliance were used to estimate the overall benefit/risk ratio of calcium channel blocker treatment in this double blind controlled study. Except for symptoms of dizziness, adverse effects were generally mild and did not lead to patient withdrawal. During the trial two patients died, one after 4 and one after 15 days of treatment with mibefradil 50 mg. The benefit/risk ratio was

assessed shows that, according to this approach, diltiazem low dose performed significantly better than did low dose mibefradil, and tended to perform better than did low dose amlodipine.

	Amlodipine		*Diltiazem*		*Mibefradil*	
	5 mg	*10 mg*	*200 mg*	*300 mg*	*50 mg*	*100 mg*
Improved exercise tolerance	7.5^{xxx}	9.9^{xxx}	9.6^{xx}	12.4	16.0	20.8
95% CI	−7.5, 22.5	−9.1, 29.1	−10.9, 30.1	−9.9, 34.9	0.0, 32.0	2.3, 38.8
Reduced patient compliance	3.5	4.5	0.9^{x}	3.0^{x}	7.3	10.8
95% CI	0.9, 8.9	1.5, 10.4	0, 5.0	0.8, 8.7	3.2, 13.9	5.8, 18.3
Ratio improved exercise tolerance/reduced patient compliance	$2.1^{(x)}$	2.2	10.7^{x}	4.1	2.2	1.9
95% CI	0.4, 7.2	0.4, 7.3	5.7, 18.3	1.4, 10.1	0.4, 7.5	0.2, 6.4

ETT = exercise testing, 95% CI = 95% confidence intervals. $^{(x)}0.05 < P < 0.1$ *vs* diltiazem comparable dose. $^{x}P < 0.05$ *vs* mibefradil comparable dose. $^{xx}P < 0.02$ *vs* mibefradil comparable dose. $^{xxx}P < 0.001$ *vs* mibefradil comparable dose.

4 Computing the Confidence Intervals of the Ratio of Normal Variables with the Quadratic Method

The computation of confidence intervals of ratios of normal variables is not straightforward, and requires a special approach, called the quadratic method. Using this method, the confidence intervals of ratios can be approximated as follows.

For the normal variables y and x, call their ratio r.

$$r = y/x$$

The equation can be rewritten.

$$y = rx$$

Consider the term y-rx. With normal data this term can be written as a linear composite term.

$d = y - rx$, where d must have a mean of zero and its variance s^2 can be written as:

$$s^2 = {s_y}^2 - 2\,r\,\mathrm{Cov}\,(xy) + r\,{s_x}^2,$$

where Cov (xy) = the covariance of the two variables = $SP_{xy}/SS_x\ SS_y$.

(SP = sum of products, SS = sum of squares).

Because r = normally distributed the t-statistic t can be used for computing its variance:

$$t^2 = (y - rx)^2/{s_y}^2 - 2\,r\mathrm{Cov}(xy) + r^2{s_x}^2).$$

The above equation can be rewritten as a quadratic equation $p = aq^2 + bq + c$, and the quadratic formula $f = -b \pm \sqrt{(b^2 - 4ac)}$ can be used for finding the upper and lower values of the confidence interval.

$$\text{quadratic formula } f = 1 - t^2 s_x{}^2 / x^2$$

$$\text{center of confidence interval } C = \left(y/x - t^2 \text{Cov}\,(xy)/x^2\right)/f$$

$$\text{standard error} = \left\{ s_y{}^2 - 2\, y/x\, \text{Cov}\,(xy) + y^2/x^2 s_x{}^2 - t\, s_x{}^2/x^2 \left[s_y{}^2 - \text{Cov}\,(xy)^2/s_x{}^2 \right] \right\}^{1/2}/(xf)$$

$$\text{confidence interval} = C \pm t \text{ times standard error.}$$

The above computations may be pretty complex for non-mathematicians, but additional explanations can be found in the literature, for example in Dunlap and Silver (in: Confidence intervals and standard errors for ratios of normal variables, Behavior Research Methods, Instruments & Computers 1986; 18: 469–71). Also, confidence interval calculators for proportions and even for comparisons of proportions are available on the Internet, for example at www.medcalc.org/calc.comparison_of_proportions (MedCalc Software bvba, Ostend Belgium). We should add, that in the example given the benefit/risk ratios are paired, and should, therefore, be adjusted for levels of correlation. For the purpose the covariance terms are included in the above equations.

5 Discussion

Benefit risk assessment may be the basis for drug administrations to monitor ongoing drug safety data. Despite the draft for an implementation plan in 2018, a document of over 7000 words emphasizing the need for enhanced and better communicated assessments and standardized operating procedures for identifying benefits outweighing risks, and providing extensive evidence about drug's safety, the entire document makes no quantitative claims, and states, in its final lines, that the benefit risk framework the FDA works with, is no more than a structured *qualitative* approach. Nonetheless it is promised, that systematic ways for more quantitative benefit risk analyses will be explored. The current chapter gives an example of how, with the help of for example the quadratic method confidence intervals of the ratio of normal variables like those of efficacy and safety scores can be readily computed. It may be a first step to a more quantitative approach to the analysis of drug safety. As an example a trial performed by our group (Br J Clin Pharmacol (1991; 50: 545–60) was used as an example.

6 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 9
Equivalence, Inferiority and Superiority Testing of Adverse Effects

Abstract Instead of testing a null hypothesis of no adverse effect, the presence of adverse effect can also be assessed with confidence intervals, and confirmed if a priori defined boundaries meet those confidence intervals. Boundaries commonly set are those of.

equivalencce,
inferiority,
superiority.

This chapter assesses how testing works and why and when it is particularly relevant.

If, in a controlled trial a new treatment is compared to a standard treatment instead of placebo, there is a big risk of finding small differences, and testing against prior boundaries is particularly helpful.

Keywords Null hypothesis · Adverse effect · Confidence intervals · Priori defined boundaries · Equivalencce · Inferiority · Superiority · Comparison versus placebo

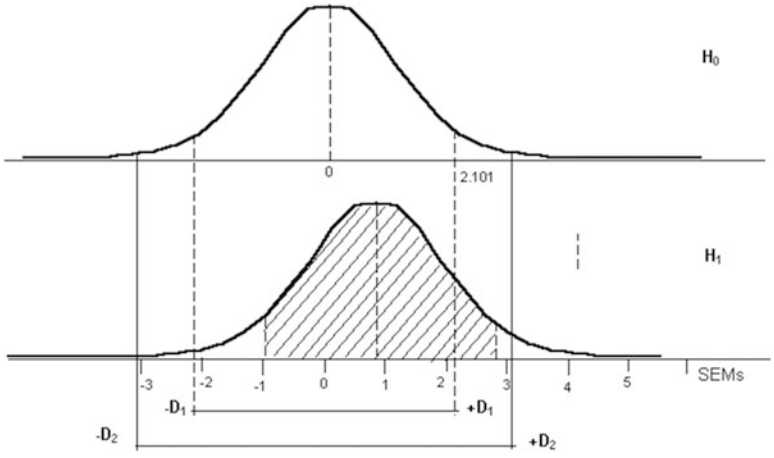
1 Introduction

Equivalence, inferiority and superiority testing instead of null hypothesis testing is increasingly common in clinical research. 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 be unethical. In the situation a new treatment must be compared with standard treatment. The latter comparison is at risk of finding little difference. This is true not only with efficacy but also with safety data. Therefore, just like with efficacy data, instead of testing a null hypothesis of no adverse effect, the presence of adverse effect can also be assessed with confidence intervals, and confirmed if a priori defined boundaries meet those confidence intervals.

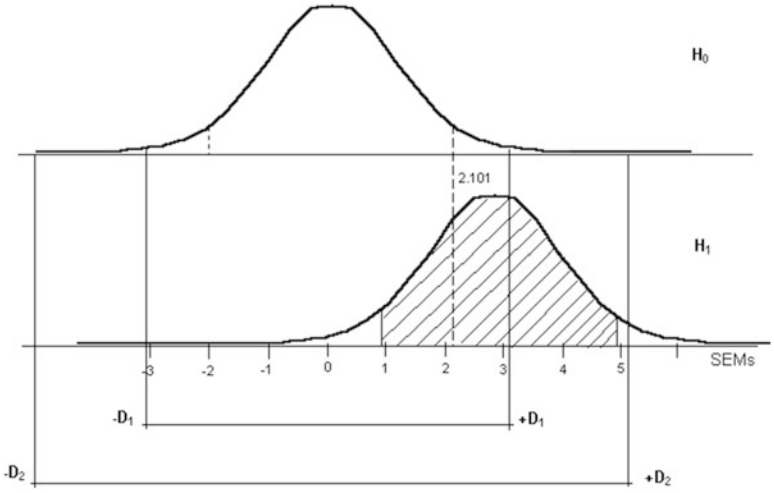
T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_9

2 How Does Traditional Equivalence, Inferiority and Superiority Testing Work

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 difference. The underneath graph 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.



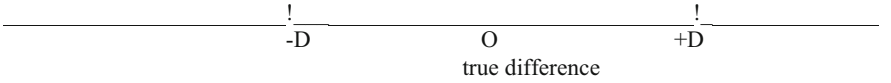
What we observe is, that the mean of our trial is only 0.9 standard errors of the mean (SEMs) distant from 0. which is far too little to reject the null-hypothesis. Our result is not significantly different from 0. Whether our result is equivalent to 0, depends on our prior defined criterium of equivalence. In the above 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 means, 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% confidence intervals (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.



Another example is given in the above graph. Here the mean result of our trial is larger: mean value of study 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 a significant difference, and equivalence.

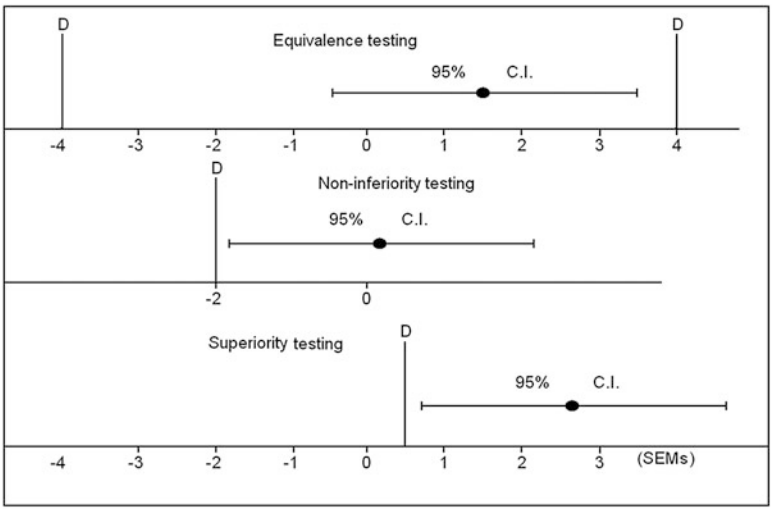
The underneath table shows, that any confidence interval (95% confidence intervals (CIs) intervals are indicated between the brackets in each of the examples), that does not overlap zero, is statistically different from zero. Only intervals between the pre-specified range of equivalence –D to + D present 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, a significant difference from zero. Situation 8 presents neither significant difference, nor 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 >



Testing equivalence of two treatments is different from testing their difference. We will here 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 will be demonstrated, if not, there will be room for uncertainty. The above table shows the discrepancies between significance testing 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 roles 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.

Like with equivalence testing, inferiority and superiority testing uses 95% confidence intervals of the outcome data and prespecified ranges demonstrating inferiority/superiority or not. In the underneath graphs additional examples are given.



With equivalence or non-inferiority/superiority testing we have prior arguments to assume clinically relevant intervals of the levels of similarity, non-inferiority etc. of the new treatment versus control, unlike a statistically significant difference between the new treatment and control. Boundaries of equivalence or non-inferiority are 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 will be accepted.

3 Why Equivalence, Inferiority and Superiority Testing of Adverse Effects

Equivalence testing is particularly relevant, if, in a controlled study, a new treatment is compared to a standard treatment instead of placebo. This is, because there is a big risk of finding a small difference, and equivalence testing is a meaningful way for data assessment not able to reject a null hypothesis due to small differences. This may also be true with adverse effects. Instead of testing a null hypothesis of no adverse effect, the presence of adverse effects can be assessed with 95% confidence intervals, and can be confirmed, if a priori defined boundaries meet those 95% confidence intervals. The only difference between equivalence and inferiority/superiority testing is the use of a single instead of two boundaries. In the next section examples will be given. The numbers of patients with dizziness, palpitations, and nasal congestion after treatment will be assessed, both in the treatment, and in the control group. Odds ratios (ORs) of and their logarithmic transformations will be used for estimating the confidence intervals of the presence of adverse effects in the treatment versus the control group.

4 Example 1

The presence of dizziness after alpha blocker and beta blocker was assessed in 310 patients with Raynaud's phenomenon.

	alpha blocker	beta blocker
yes	80 a	100 b
no	80 c	50 d

The range of equivalence was a priori defined between an odds ratio of 0.5 and 2.0, meaning about half to twice as many cases in the alpha blocker group as those in the beta blocker group.

Odds ratio (OR) from the above data	= (80/80) / (100/50) = 0.5
the standard error of the log OR	= √ (1/a + 1/b + 1/c + 1/d) = √ (1/80 + 1/80 + 1/100 + 1/50) = √ (0.0125 +..... + 0.02) = √ (0.055) = 0.2345
log OR	= log 0.5 = -0.6931
95% confidence interval of log OR	= -0.6931 ± 2 x 0.2345 = -0.6931 ± 0.4690 = between -1.1621 and 0.2241
95% confidence interval of OR	= the antilogs of the above = between 0.3128 and 1.2512.

The above 95% confidence interval is not entirely but only partly within the range of equivalence, and, so, the presence of equivalence of the odds in the treatment and control group can not be confirmed. It is uncertain, because only the right end of the 95% confidence interval is within the range of equivalence, and the left end is not so.

5 Example 2

The presence of palpitations after alpha blocker and beta blocker was assessed in 310 patients with Raynaud's phenomenon.

	alpha blocker	beta blocker
yes	50 a	20 b
no	110 c	130 d

The range of inferiority was a priori defined as an odds ratio > 3.0, meaning about. three times as many cases in the alpha blocker group as those in the beta blocker group.

Odds ratio (OR) from the above data	= (50/110) / (20/130) = 2.9545
the standard error of the log OR	= √ (1/a + 1/b + 1/c + 1/d) = √ (1/50 + 1/110 + 1/20 + 1/130) = √ (0.02 +..... + 0.00769) = √ (0.08669) = 0.2944
log OR	= log 2.9545 = 1.0833
95% confidence interval of log OR	= 1.0833 ± 2 x 0.2944 = 1.0833 ± 0.5889 = between 0.4944 and 1.6722
95% confidence interval of OR	= the antilogs (2ndf or shift log) of the above = between 1.6400 and 5.3239.

The above 95% confidence interval is not entirely but only partly within the range of inferiority, and, so, the presence of inferiority of the odds in the treatment and control group can not be confirmed. It is not entirely without either, and, so, the result of the assessment is uncertain.

6 Example 3

The presence of nasal congestion after alpha blocker and beta blocker was assessed in 310 patients with Raynaud's phenomenon.

	alpha blocker	beta blocker
yes	100 a	100 b
no	60 c	50 d

The range of superiority was a priori defined as an odds ratio > 0.5, meaning about half as few cases in the alpha blocker group as those in the beta blocker group.

Odds ratio (OR) from the above data	= (100/60) / (100/50) = 0.8333
the standard error of the log OR	= √ (1/a + 1/b + 1/c + 1/d) = √ (1/100 + 1/60 + 1/1000 + 1/50) = √ (0.01 +..... + 0.02) = √ (0.05667) = 0.2381
log OR	= log 0.8333 = -0.1824
95% confidence interval of log OR	= -0.1824 ± 2 x 0.2381

The above 95% confidence interval is entirely within the range of superiority, and, so, the presence of superiority of the odds in the treatment and control group can be confirmed. The alpha blocker performs better than the beta blocker with respect to the adverse effect nasal congestion.

7 Discussion

Equivalence, inferiority and superiority testing instead of null hypothesis testing is increasingly common in clinical research. A study unable to find a difference is not the same as an equivalent study. 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. The same is true with adverse effect assessments in clinical trials. Therefore, instead of testing a null hypothesis of no adverse effect, the presence of adverse effect can also be assessed with confidence intervals, and confirmed if it meets a prior defined boundaries.

8 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,

Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been written by the same authors, and they have been edited by Springer Heidelberg Germany.

Part II
The Analysis of Dependent Adverse Effects

Chapter 10
Independent and Dependent Adverse Effects

Abstract Adverse effects may be either dependent or independent of the main outcome. An adverse effect of alpha blockers is dizziness, and this occurs independently of the main outcome "alleviation of Raynaud's phenomenon". In contrast, the adverse effect "increased calorie intake" occurs with "increased exercise", and this adverse effect is very dependent on the main outcome "weight loss". Random heterogeneities, outliers, confounders, interaction factors are common in clinical trials, and all of them can be considered as kinds of adverse effects of the dependent type. Random regressions and random analyses of variance, high dimensional clustering, Bayesian methods are helpful for their analysis and many more methods are possible.

A dependent adverse effect must be significantly related not only to the intervention but also the outcome.

This chapter particularly addresses causal relationships as the underlying mechanism of a dependent adverse effect. Path statistics, path analysis, partial correlations, including d-separations, partial correlation analysis, and higher order partial correlations are particularly suitable for the purpose, but additional methods do exist.

Keywords Dependent adverse effect · Random heterogeneities · Outliers · Confounders · Interaction factors · Random regressions · Random analyses of variance · High dimensional clustering · Bayesian methods · Causal relationships · Path statistics · Path analysis · Partial correlations · D-separations · Partial correlation analysis · Higher order partial correlations

1 Introduction

The Chaps. 2, 3, 4, 5, 6, 7, 8 and 9 only assessed independent adverse effects. It is time, that we addressed the issue of *dependency* of adverse effects. Adverse effects may be either dependent or independent of the main outcome. For example, an adverse effect of alpha blockers is dizziness, and this occurs independently of the main outcome "alleviation of Raynaud's phenomenon". In contrast, the adverse

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_10

effect “increased calorie intake” occurs with “increased exercise”, and this adverse effect is very dependent on the main outcome “weight loss”. Random heterogeneities, outliers, confounders, interaction factors are common in clinical trials, and all of them can be considered as kinds of adverse effects of the dependent type. Random regressions and random analyses of variance, high dimensional clustering, Bayesian methods are helpful for their analysis and many more methods are possible. Causal relationships are hard to prove, and are, sometimes, called the deepest enigma of mankind. Particularly, Bayesian methods are helpful for supporting causal relationships. This is because it uses path analysis. Path analysis takes add-up sums of standardized regression coefficients for better estimation of multiple step relationships. Multiple steps provide more sensitive predictions than a single step does, that is, if they do statistically significantly so. A series of significantly positive-related events interpreted as a possibly causal pathway, will be more easily believed to be truly causal, than the results of multiple independent predictors in a multiple regression analysis will. The latter only demonstrates associations, and associations may very well be due to time effects and place effects and other confounders, and need not necessarily be causal. In order to assess whether a dependent adverse effect is significantly dependent on the outcome or not, special methods are used based on the underlying mechanisms of dependency, including causal, pharmacological, interaction, subgroup, carryover, pleiotropic drug mechanisms, and more, may be responsible for dependent adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to the intervention but also the outcome. This chapter will particularly address causal relationships as the underlying mechanism of a dependent adverse effect. Path statistics, path analysis, partial correlations, including d-separations, partial correlation analysis, and higher order partial correlations are particularly suitable for the purpose, but additional methods do exist. We will describe step by step analyses of data files available through extras.springer.com in order for readers to be able to rehearse statistical analyses for themselves.

2 Multiple Path Analysis

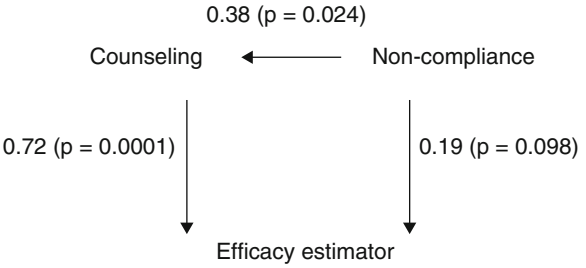
As an example, “non-compliance with treatment” affects the efficacy estimator (outcome). Therefore, in the protocol counseling was included. We assumed, that, the more non-compliance, the better the effect of counseling on the efficacy outcome of the study would be, and, thus, that counseling would be a dependent adverse effect of non-compliance on the efficacy estimator.

Path statistics uses add-up regression coefficients for better estimation of multiple step relationships. Because regression coefficients have the same unit as their variable, they can not be added up, unless multiplied with the ratio of their standard deviations. Standardized regression coefficients are, otherwise, called path statistics. The underneath figure gives a path diagram of variables. We should add, that

standardized B-values (standardized regression coefficients) are routinely provided in the output of regression analyses as given by most statistical software programs. The relationship between standardized and non-standardized B-values is given underneath:

$$\text{Standardized B} = \text{non-standardized B x } SD_x/SD_y.$$

SD = standard deviation, B = regression coefficient. A convenient property of the standardized B is, that it has a variance of 1, and that its unit is not grams, mmols, mms etc. but rather SE (standard error) units. Standardized B-values can, therefore, be added, subtracted, multiplied etc.



In the above table the arrows are the paths, and standardized regression coefficients are added to the arrows. The graph shows a structural equation model. Single path analysis gives a standardized regression coefficient of 0.19. This underestimates the real effect of non-compliance on the efficacy estimator. Two step path analysis is more realistic, and shows, that the add-up path statistic is larger, and equals

$$0.19 + 0.38 \times 0.72 = 0.46.$$

The two-path statistic of 0.46 is, thus, a lot better than the single path statistic of 0.19 with an increase of 60%.

It is easy to demonstrate, that multiple steps provide a more sensitive prediction than a single step model does, at least, if they are statistically significant. And multiple paths similarly do so as compared to single paths. A series of significantly positive-related events interpreted as a possibly causal pathway, will be more easily believed to be truly causal, than the results of multiple independent predictors in a multiple regression analysis will.

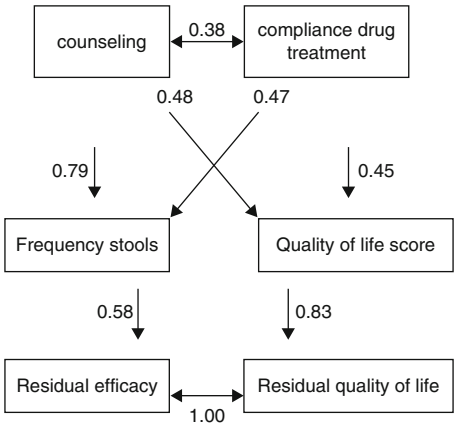
The above procedure shows a significant relationship between all of the factors. Particularly counseling does have a significant negative correlation with non-compliance and positive correlation with the efficacy estimator. This means, that with $p < 0.10$ as cut-off for statistical significance, counseling is, indeed, a significant dependent adverse effect of non-compliance on the efficacy estimator (the outcome).

The above example will be used once more, but now quality of life scores will be added as an additional variable to the model. The data file is in extras.springer.com, and is entitled "pathstatistics". A 35 patient data file summarizes between-variable relationships. The first 10 patient data are underneath.

var. 1 = stools per month
var. 2 = qol (quality of life score)
var. 3 = counselings per month
var. 4 = compliance = non-compliances with drug treatment

Var 1	2	3	4
24,00	69,00	8,00	25,00
30,00	110,00	13,00	30,00
25,00	78,00	15,00	25,00
35,00	103,00	10,00	31,00
39,00	103,00	9,00	36,00
30,00	102,00	10,00	33,00
27,00	76,00	8,00	22,00
14,00	75,00	5,00	18,00
39,00	99,00	13,00	14,00
42,00	107,00	15,00	30,00

The linear correlation coefficients between all of the variables are computed. They are equal to the standardized regression coefficients and are also called path statistics. Their units are SE (standard error) units, and they can, therefore, conveniently be added up or subtracted, unlike usual regression coefficients that have the same units as those of the variables they stem from like mmol/l, mg, sec etc.



The above graph shows a structural equation model with 8 instead of 3 different path statistics computed as explained above. A nice thing about path statistics is that they can be multiplied in order to compute indirect effects of one variable through another variable on a subsequent variable. For example:

effect of frequency of stools through counseling on quality of life score
$0.79 \times 0.48 = 0.38$
effect of frequency of stools through non-compliance on quality of life score
$0.45 \times 0.47 = 0.21$.

As path statistics can be multiplied with one another, they can be used for assessing multiple step regressions, and even, as a nonmathematical method, for performing complex multivariate regressions, the traditional mathematical alternative of which is much more complex. They can also be used for testing several "null hypotheses of no effect" of no dependent adverse effects, for example, counseling on frequency of stools and on quality of life.

3 Partial Correlations

Causality in the above structural equation models is supported even more with the help of d-separations using partial correlation modeling. D-separations literally means separations based on dependencies. With d-separations the likelihood distribution of a predictor on an outcome is estimated on the condition that another predictor is held constant. Bayesian networks, otherwise called DAGs (directed acyclic graphs), are here used as a way to figure out causal structures of biological models, such, that no experiments are needed. For example, the underneath DAG shows that blood sugar predicts stomach acidity, and stomach acidity predicts hunger.

$$\text{blood sugar} \rightarrow \text{stomach acidity} \rightarrow \text{hunger}$$

The above simple DAG model asserts, that blood sugar causes stomach acidity directly, and, that stomach acidity causes hunger directly. This model implies, that blood sugar and hunger are correlated. If acidity, being held constant, causes the significant correlation between blood sugar and hunger not to disappear, then the originally established correlation between blood sugar and hunger must have been causal. Even with multiple predictors, partial correlations analysis is possible. An example will be given underneath. Only the first 10 patients are in the underneath table.

Var 1 weightloss	Var 2 exercise	Var 3 calorieintake	Var 4 interaction	Var 5 age
1,00	0,00	1000,00	0,00	45,00
29,00	0,00	1000,00	0,00	53,00
2,00	0,00	3000,00	0,00	64,00
1,00	0,00	3000,00	0,00	64,00
28,00	6,00	3000,00	18000,00	34,00
27,00	6,00	3000,00	18000,00	25,00
30,00	6,00	3000,00	18000,00	34,00
27,00	6,00	1000,00	6000,00	45,00
29,00	0,00	2000,00	0,00	52,00
31,00	3,00	2000,00	6000,00	59,00

The entire data file gives the data of a simulated 64 patient study of the effects of exercise on weight loss with calorie intake as covariate. For convenience the data file is stored at extras.springer.com, and is entitled "partialcorrelations". We wish to perform a multiple linear regression of these data with weight loss as dependent (y) and exercise (x_1) and calorie intake (x_2) as independent predictor variables. Because the independent variables should not correlate too strong, first a correlation matrix is calculated as given below.

Correlations

		weightloss	exercise	calorieintake
weightloss	Pearson Correlation	1	,405**	-,304*
	Sig. (2-tailed)		,001	,015
	N	64	64	64
exercise	Pearson Correlation	,405**	1	,390**
	Sig. (2-tailed)	,001		,001
	N	64	64	64
calorieintake	Pearson Correlation	-,304*	,390**	1
	Sig. (2-tailed)	,015	,001	
	N	64	64	64

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation coefficients >0.80 or < −0.80 indicate collinearity, and indicate, that multiple regression is not valid. This is however, not so, and we can, thus, proceed.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	34,279	2,651		12,930	,000
	interaction	,001	,000	,868	3,183	,002
	exercise	-,238	,966	-,058	-,246	,807
	calorieintake	-,009	,002	-,813	-6,240	,000

a. Dependent Variable: weightloss

The above table gives the results of the multiple linear regression. Both calorie intake and exercise are significant independent predictors of weight loss in the univariate models. However, exercise makes you hungry and patients on weight training may be inclined to reduce (or increase) their calorie intake. So, the presence of an interaction between calorie intake and exercise on weight loss is very well possible. In order to check this, an interaction variable (x_3 = calorie intake * exercise, with * symbol of multiplication) was added to the model. After the addition of the interaction variable to the regression model, exercise is no longer significant and interaction on the outcome is significant at $p = 0.002$. There is, obviously, interaction in the study, and the overall analysis of the data is no longer relevant. In order to find the best method to identify the true effect of exercise, the study should be repeated with calorie intake held constant. Instead of this laborious exercise, a partial correlation analysis with calorie intake held constant can be adequately performed, and would provide virtually the same result.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,405[a]	,164	,151	9,73224

a. Predictors: (Constant), exercise

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,644[a]	,415	,396	8,20777

a. Predictors: (Constant), calorieintake, exercise

The above tables show, that the simple linear regression between exercise and weight loss produced a correlation coefficient (r-value) of 0.405, and the multiple correlation coefficient, including calorie intake as additional predictor, was larger,

0.644. The r-square values are often interpreted as the % certainty about the outcome given by the regression analysis, and we can observe from the example that it rises from 0.164 to 0.415, meaning that, with two predictors, we have 42% instead of 16% certainty. The addition of the second independent variable provided 26% more certainty. But, what about the correlation between the second variable (x_2) and the outcome (y). Is it equal to the square root of 0.26 = 0.51 (51%)? No, this is not so, because multiple regression is a method that finds the best fit model for all of the data, and because, if you add new data, then all of the previously calculated relationships will change. The change in y caused by the addition of a second variable can be calculated by removing the amount of certainty provided by the presence of a novel variable.

$$\text{Novel y values} = \text{y values} - \text{mean y} - r_{\text{y vs x2}}\left(SD_y/SD_{x2}\right).$$

SD = standard deviation, vs = versus. Similarly the novel x_1 values can be calculated. Once this has been done for all individuals, an ordinary correlation between the novel values can be calculated. The novel correlation is interpreted as the correlation between y en x_1 with x_2 held constant, and is, otherwise, called the partial correlation.

With additional x variables, even higher-order partial correlations can be calculated, which are computationally very intensive, but calculations are pretty much the same. The interpretation is straightforward, the partial correlation between y en x_1 with two additional x variables is the correlation between y en x_1 with x_2 and x_3 held constant. What is the clinical relevance of partial correlations. It removes the effects of interactions of predictor variables on the outcome variable, and, thus, establishes what would have happened, if there had been no interaction. Also, it provides support for a causal relationship between variables: if, with 3 paired variables, one of three is held constant, and if this causes the previously significant correlation between the other two not to disappear, then the originally established correlation between the latter two will probably be causal. For the current purpose d-separations are also relevant.

They can be used for testing the "null hypotheses of no effect" (of no dependent adverse effects, with, for example, calorie intake as a dependent adverse effect of physical exercise on weight loss).

In addition to d-separations also a partial correlation analysis can be performed for testing the presence of dependent adverse effects of the causal type, using SPSS, menu module Correlations. The data file called "partialcorrelations", available in extras.springer.com is used once more. It is opened it in your computer mounted with SPSS statistical software.

Command

Analyze. . . .Correlate. . . .Partial. . . .Variables: enter weight loss and exercise. . . . Controlling for: enter calorie intake. . . .click OK.

Correlations

Control Variables			weightloss	exercise
calorieintake	weightloss	Correlation	1,000	,596
		Significance (2-tailed)	.	,000
		df	0	61
	exercise	Correlation	,596	1,000
		Significance (2-tailed)	,000	.
		df	61	0

The above table shows that, with calorie intake held constant, exercise is a significant positive predictor of weight loss with a correlation coefficient of 0.596 and a p-value of 0.0001. It is interesting to observe that the partial correlation coefficient between weight loss and exercise is much larger than the simple correlation coefficient between weight loss and exercise (correlation coefficient = 0.405, former table). Why do we not have to account interaction with partial correlations. This is simply because, if you hold a predictor fixed, this fixed predictor can no longer change and interact in a multiple regression model. In this way partial correlations demonstrates the presence of a causal relationship between the assumed dependent adverse effect and an outcome.

4 Higher Order Partial Correlations

Instead of a single variable, also multiple variables can be held constant in higher order partial correlation analyses. Age may affect all of the three variables already in the model. The effect of exercise on weight loss with calorie intake and age fixed is shown. The correlation coefficient is still very significant as shown in the underneath table.

Correlations

Control Variables			weightloss	exercise
age & calorieintake	weightloss	Correlation	1,000	,541
		Significance (2-tailed)	.	,000
		df	0	60
	exercise	Correlation	,541	1,000
		Significance (2-tailed)	,000	.
		df	60	0

In the above partial correlations models the correlations between exercise and weight loss did not vanish. In contrast, the levels of correlation even increased as compared to the level of correlation in the univariate model of weight loss versus

exercise. This would mean that the concept of causality is supported. A causal relationship, and, thus, the presence of calorieintake and age as dependent adverse effects is supported.

5 Bayesian Networks, Pleiotropy Research

The statistical proof of dependent adverse effects of the causal type was the main subject of this chapter. Bayesian structural equation modeling uses path analyses, and is particularly successful for the purpose. Path analysis uses add-up sums and multiplications of standardized regression coefficients for better estimation of multiple step relationships. Multiple steps provide a more sensitive prediction than a single step do, particularly if they do significantly so. And multiple paths similarly do so, as compared to single paths. A series of significantly positive-related events interpreted as possibly causal pathways, will be more easily believed, than the results of multiple independent predictors in a multiple regression analysis will be.

Bayes theorem, although, traditionally, used for analyzing qualitative diagnostic tests, is currently mostly applied for computation of Bayes factors. A Bayes factor is the ratio of two likelihood distributions, one of your data (posterior distribution), and one of relevant historical data (prior distribution).

The equality:

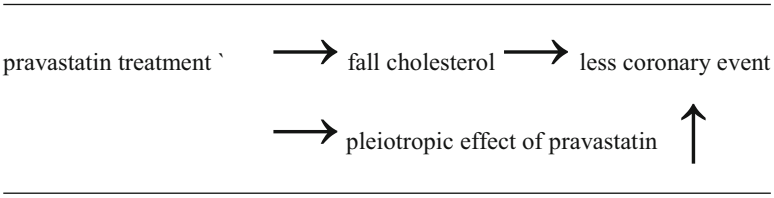
"**prior likelihood distribution x Bayes factor = posterior likelihood distribution**"

looks much like the equalities used with path analysis, and this is not only textually so but also conceptionally. An example is given.

"[**path statistic 1**] x [**path statistic 2**] = [**effect of Var 1 through Var 2 on Var 3**]",

where Var = variable.

The similarity is so much so, that, currently, stepwise path analyses are increasingly named Bayesian networks, particularly, if they are used for causality research, and, if they include multiple rather than single paths. Another example is given in the graph of a kind of structural equation model underneath. All effects of pravastatin are given. Three horizontal arrows indicate a positive significant correlation to the main outcome less coronary events. It shows a pleiotropic mechanism as dependent adverse effect on the study outcome.



the above effect of randomized treatment with pravastatine on the risk of coronary events depended not only on the amount of cholesterol change but also, directly, of the effect of the randomized treatment with pravastatin on the final outcome. This latter effect can be interpreted as a pleiotropic adverse effect of pravastatin, on the outcome les coronary events which is different from the main effects of pravastatin through cholesterol.

6 Discussion

In this chapter are examples of path analyses. With multiple variables they are called Bayesian networks, which are networks with a particular eye towards causality. Causality is even better substructured with partial correlation analysis of indirect predictors in a Bayesian network. Terms like structural equation modeling and DAGs (directed acyclic graphs) are commonly applied within this context. Path analyses are successfully applied for testing a hypothesis of adverse effects due to causal mechanisms. For those not fond of complex statistical reasoning, the above text may be somewhat sophisticated. It may then be of help to focus on partial correlation as an adverse effect due to a causal mechanism.

7 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 11
Categorical Predictors Assessed as Dependent Adverse Effects

Abstract In this chapter examples are given of clinical studies, where the adverse effect has a categorical, rather than continuous pattern. The presence of the adverse effect can be easily missed, if its categorical pattern is analyzed as a linear variable. The linearly structured variable race was such an adverse effect of the effects of age and gender on an outcome like physical strength score. Also, the linearly structured numbers of concomitant medications was such an adverse effect of the effect of age on the risk of iatrogenic admissions to hospital.

A linear or logistic regression with categories assessed as separate independent variables instead of a single continuous variable is adequate for analysis, and, in the examples given in this chapter, it was able to demonstrate the presence of dependent adverse effects in the form of significant and insignificant predictor categories on the outcome.

Keywords Dependent adverse effect · Categorical data · Continuous data · Linear regression · Categorical predictor · Significant and insignificant predictor categories

1 Introduction

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. A dependent adverse effect must be significantly related not only to the intervention but also the outcome. In this chapter examples will be given where the adverse effect has a categorical rather than continuous pattern. It is shown, that the presence of the adverse effect can be easily missed, if its categorical pattern is analyzed as a linear variable. The linearly structured variable race may be such an adverse effect of, for example, the effects of age and gender on an outcome like physical strength score. Also, the linearly structured numbers of concomitant medications may be such an adverse effect of, for example, the effect of age on the risk of iatrogenic admissions to hospital.

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_11

2 Example 1

In a study of 60 patients with different races, the effects of age and genders on the outcome physical strength were assessed. The presence of one of four races, namely white, hispanic, black and asian was available in the data file. Because whites and blacks were thought to have better physical strength than hispanic and asians the variable race was included in the analysis as a possible adverse effect of the two predictors on the outcome. A variable like race, may or may not have an incremental function, and, therefore, linear regression may or may not be appropriate for assessing their effect on the outcome.

The effects on physical strength (scores 0–100) were assessed in 60 subjects of different races (hispanics (1), blacks (2), asians (3),and whites (4)), ages (years), and genders (0 = female, 1 = male). The first 10 patients are in the table underneath.

patient number	physical strength	race	age	gender
1	70,00	1,00	35,00	1,00
2	77,00	1,00	55,00	0,00
3	66,00	1,00	70,00	1,00
4	59,00	1,00	55,00	0,00
5	71,00	1,00	45,00	1,00
6	72,00	1,00	47,00	1,00
7	45,00	1,00	75,00	0,00
8	85,00	1,00	83,00	1,00
9	70,00	1,00	35,00	1,00
10	77,00	1,00	49,00	1,00

The entire data file is in extras.springer.com, and is entitled “categorical”. Start by opening the data file in your computer with SPSS installed.

Command

Analyze. . . .Regression. . . .Linear. . . .Dependent: physical strength score. . . . Independent: race, age, gender. . . .OK.

The table in the output sheets shows, that age and gender are significant predictors but race is not.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	79,528	8,657		9,186	,000
	race	,511	1,454	,042	,351	,727
	age	-,242	,117	-,260	-2,071	,043
	gender	9,575	3,417	,349	2,802	,007

a. Dependent Variable: strengthscore

The variable race is analyzed as a stepwise rising function from 1 to 4, and the linear regression model assumes that the outcome variable will rise (or fall) simultaneously and linearly, but this needs not be necessarily so. Next, a categorical analysis will be performed. For that purpose we first need to restructure the data file replacing the race single race variable with the values 1–4 with four separate dichotomous values with the values 0,00 or 1,00 as demonstrated underneath.

Command
click race. . . .click Edit. . . .click Copy. . . .click a new "var". . . .click Paste. . . .highlight the values 2–4. . . .delete and replace with 0,00 values. . . .perform the same procedure subsequently for the other races.

In the data screen the underneath restructured pattern is given.

patient number	physical strength whites	race	age	gender	race 1 hispanics	race 2	race 3 blacks	race 4 asians
1	70,00	1,00	35,00	1,00	1,00	0,00	0,00	0,00
2	77,00	1,00	55,00	0,00	1,00	0,00	0,00	0,00
3	66,00	1,00	70,00	1,00	1,00	0,00	0,00	0,00
4	59,00	1,00	55,00	0,00	1,00	0,00	0,00	0,00
5	71,00	1,00	45,00	1,00	1,00	0,00	0,00	0,00
6	72,00	1,00	47,00	1,00	1,00	0,00	0,00	0,00
7	45,00	1,00	75,00	0,00	1,00	0,00	0,00	0,00
8	85,00	1,00	83,00	1,00	1,00	0,00	0,00	0,00
9	70,00	1,00	35,00	1,00	1,00	0,00	0,00	0,00
10	77,00	1,00	49,00	1,00	1,00	0,00	0,00	0,00
.....								
....								

For convenience the restructure data file is also given in extras.springer.com and is entitled "categoricalrestructured". For a categorical analysis the above commands are given once more, but now the independent variables are entered slightly differently.

Command
Analyze. . . .Regression. . . .Linear. . . .Dependent: physical strength score. . . . Independent: race 2, race 3, race 4, age, gender. . . .click OK.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	72,650	5,528		13,143	,000
	race2	17,424	3,074	,559	5,668	,000
	race3	-6,286	3,141	-,202	-2,001	,050
	race4	9,661	3,166	,310	3,051	,004
	age	-,140	,081	-,150	-1,716	,092
	gender	5,893	2,403	,215	2,452	,017

a. Dependent Variable: strengthscore

The above table shows the outcome sheets. Races 2–4 are now significant predictors of physical strength. Obviously, compared to the presence of the hispanic race, the black and white races are significant positive predictors of physical strength ($p = 0.0001$ and 0.004 respectively), the asian race is a significant negative predictor ($p = 0.050$). And race in general can thus be interpreted as a significant adverse effect with a categorical rather than linear pattern. The categorical analysis allows to conclude that race is, indeed, a dependent adverse effect of the effect of age and gender on physical strength, that would have been missed if a linear instead of categorical analysis would have been performed.

3 Example 2

Numbers of concomitant medications (co-medications) may be an adverse effect of the effects of age and gender on the risk of iatrogenic admission to hospital. Numbers of co-medications may be positively correlated with risk of admission to hospital due to adverse drug effects. In this example we will use the data from a recently published 2000 patient cohort study from our group (Atiqi et al., Int J Clin Pharmacol and Ther 2010: 48; 517–25) about adverse-drug-effect-admissions. An SPSS data file is in extras.springer.com, and is entitled "iatrogenicadmissions".

A logistic regression with risk of iatrogenic admission as outcome and age (variable 1) and genders (variable 2) as predictors and numbers of co-medications and those of co-morbidities (var 10 and 9) as covariate and possibly dependent adverse effects was performed. Logodds was used as a surrogate of risk. Start by opening the data file in your computer mounted with SPSS statistical software.

Command

Analyze. . . .Regression. . . .Binary Logistic. . . .Dependent: iatrogenic hospital admission. . . . Independent: age, gender, comed, como. . . .click OK.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VAR00001	-,023	,004	32,062	1	,000	,977
	VAR00002	,089	,116	,580	1	,446	1,093
	VAR00010	,004	,072	,003	1	,953	1,004
	VAR00009	,095	,073	1,672	1	,196	1,099
	Constant	43,752	8,077	29,345	1	,000	1,003E19

a. Variable(s) entered on step 1: VAR00001, VAR00002, VAR00010, VAR00009.

VAR00001 = age
VAR00002 = gender
VAR00010 = comed (numbers of concomitant medications)
VAR00009 = como (numbers of concomitant morbidities)

In the above output sheets the performance of the predictors (var 1 and 2) and the covariates (var 10 and 9) are given. Only age (var 1) is statistically significant, and the both covariates are definitely not. And so, they should be concluded not to be dependent adverse effects.

However, the problem is, that, if scores "zero to eight" are used as a linear covariate in a logistic model, then we assume that the risk of adverse-drug effect-admissions rises linearly, but this needs not to be so. If the relationship is a stepping function, like with categories, and, if we assume a linear relationship, then we will be at risk of severely underestimating effects. In order to escape this risk, it is more appropriate to transform a quantitative estimator used as continuous variable into a categorical one. Using logistic regression in SPSS is convenient for the purpose. We need not manually transform the quantitative estimator. For the analysis we apply the usual commands.

Command

Analyze. . . .Regression. . . .Binary Logistic. . . .enter dependent variables. . . .enter independent variables. Then, open dialog box labelled categorical variables, select co-medication and transfer it into the box categorical variables, then click continue. Co-medication is now transformed into a categorical variable. Click OK. The underneath table gives the results.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VAR00001	-,024	,004	32,859	1	,000	,976
	VAR00002	,080	,117	,470	1	,493	1,084
	VAR00010			22,241	8	,004	
	VAR00010(1)	18,900	40199,059	,000	1	1,000	1,615E8
	VAR00010(2)	19,490	40199,059	,000	1	1,000	2,914E8
	VAR00010(3)	18,923	40199,059	,000	1	1,000	1,653E8
	VAR00010(4)	19,342	40199,059	,000	1	1,000	2,514E8
	VAR00010(5)	18,820	40199,059	,000	1	1,000	1,491E8
	VAR00010(6)	19,122	40199,059	,000	1	1,000	2,017E8
	VAR00010(7)	17,932	40199,059	,000	1	1,000	6,133E7
	VAR00010(8)	-1,109	56845,749	,000	1	1,000	,330
	VAR00009	,109	,076	2,047	1	,152	1,115
	Constant	25,804	40199,060	,000	1	,999	1,609E11

a. Variable(s) entered on step 1: VAR00001, VAR00002, VAR00010, VAR00009.

$$\begin{aligned}
&\mathrm{VAR00001} = \text{age}\\
&\mathrm{VAR00002} = \text{gender}\\
&\mathrm{VAR00010} = \text{comed (numbers of concomitant medications)}\\
&\mathrm{VAR00010}(1) = 1 \text{ concomitant medication}\\
&\qquad\qquad (2) = 2 \text{ concomitant medications}\\
&\qquad\qquad (3) = \ldots..\\
&\qquad\qquad (8) = 8 \text{ concomitant medications}\\
&\mathrm{VAR00009} = \text{como (numbers of concomitant morbidities)}
\end{aligned}$$

The number of co-medications (var 10) as independent variable assessed in the form of a categorical variable was, indeed, very significant at $p = 0.004$, and thus co-medications was a dependent adverse effect of the effects of age on the risk (or rather odds) of iatrogenic admission.

4 Discussion

A dependent adverse effect must be significantly related not only to the intervention but also the outcome. In this chapter examples were given where the adverse effect had a categorical rather than continuous pattern. With categorical adverse effect data the presence of the adverse effect is easily missed if its categorical pattern is analyzed

in the form of a linear pattern. Numbers of concomitant medications may be such an adverse effect of, for example, the effect of age on the risk of iatrogenic admissions to hospital. Also race may be such an adverse effect of, for example, the effects of age and gender on the outcome physical strength.

A linear or logistic regression with categories assessed as separate independent variables instead of a single continuous variable is adequate for analysis, and, in the examples, was able to demonstrate the presence of dependent adverse effects in the form of significant and insignificant predictor categories on the outcome.

5 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

1. Statistics applied to clinical studies 5th edition, 2012,
2. Machine learning in medicine a complete overview, 2015,
3. SPSS for starters and 2nd levelers 2nd edition, 2015,
4. Clinical data analysis on a pocket calculator 2nd edition, 2016,
5. Understanding clinical data analysis from published research, 2016,
6. Modern meta-analysis, 2017,
7. Regression analysis in clinical research, 2018,
8. Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 12
Adverse Effects of the Dependent Type in Crossover Trials

Abstract Carryover effect, otherwise called treatment by period interaction, is a major adverse effect of crossover studies. If the effect of a treatment carries on into the second period of treatment, then the measured response to the second period of treatment will be changed. This carryover effect can be considered an adverse effect of the dependent type. This is, because it changes the outcome of the study. We report levels of carryover required for a carryover effect to be significant, and give examples of negative crossover studies due to carryover adverse effect. If significant, then an overall analysis of the study is pretty meaningless, but it makes sense to analyze the first period of the study for treatment effect and disregard the second period.

The current chapter sows, that the adverse effect of carryover effect from the first into the second period of treatment in the control group can be easily tested for statistical significance. With a significant carryover effect, an overall treatment assessment makes no sense, but an unpaired analysis of the first periods of treatment is a worthwhile alternative for the purpose.

Keywords Carryover effect · Treatment by period interaction · Crossover studies · Adverse effect of the dependent type · Negative crossover studies due to carryover · Unpaired analysis of the first period of treatment

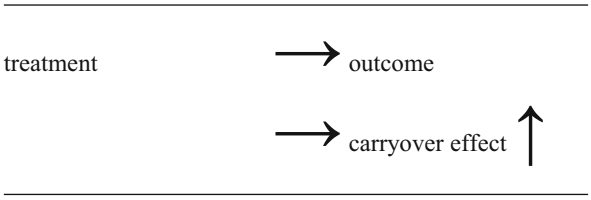
1 Introduction

In the Chaps. 10 and 11 particularly, causal and categorical predictors responsible for dependent adverse effects in clinical trials have been assessed. Many more mechanisms do exist, for example, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may be adverse effects of an intervention on an outcome. An adverse effect will be called dependent, if it is significantly related not only to intervention but also the outcome. In this chapter examples

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_12

will be given, where carryover effect from previous treatments is an adverse effect of the effect of the treatment modalities in studies on the outcome.

The underneath structure equation model shows, how it works. More information on structure equation modeling is in the Chap. 10.



Carryover effect, otherwise called treatment by period interaction, is a major adverse effect of crossover studies. If the effect of a treatment carries on into the second period of treatment, then the measured response to the second period of treatment will be changed. This carryover effect can be considered an adverse effect of the dependent type. This is, because it changes the outcome of the study. The above arrows indicate positive Pearson correlation coefficients with continuous data, but other statistics are possible, like point probabilities from Fisher exact tests with binary data. In the current chapter we will assess the presence of carryover and treatment effect in crossover studies with binary data. We will report levels of carryover effect required for a carryover effect to be significant, and give examples of negative crossover studies due to carryover adverse effect. If significant, then an overall analysis of the study is pretty meaningless, but it makes sense to analyze the first period of the study for treatment effect and disregard the second period. A level of significance will be obtained similar to that of the analysis including the carryover effect or even better.

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 underneath). If groups are symmetric, and the results are not influenced by the order of the treatments, then 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. Underneath an example is given 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. 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. In order to detect a carryover effect, we will 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 to those of the treatments B in paired comparisons. 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 Effect

We assume a 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 above). The

difference between p_C and p_A is considered to be the amount of carryover effect in the data. Fisher exact test is used for testing, whether p_C is significantly different from p_A. The 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 20–50 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 to 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 used for power calculations.

4 A Table 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), 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, $\alpha = 0.05$).

The underneath table 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. In the underneath table the computed power to demonstrate a treatment effect, in spite of the presence of a significant carryover effect, is given.

	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%).

5 A Table of Powers 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%. 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. The table below 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%. In the underneath table the power (%) to demonstrate a treatment effect, in spite of the presence of significant carryover effect, is illustrated.

	ρ	Total number of patients 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.

6 Examples

Suppose, we have a negative crossover, where probability of treatment success group II p_C, 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

$$\text{Point probability for carryover effect} = \frac{10!10!10!10!}{20!2!8!2!8!} = 0.011$$

The 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 will, 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:

$$\text{Point probability for carryover effect} = \frac{29!8!10!10!}{20!2!8!10!0!} = 0.004$$

$$\text{Cumulative tail probability} = 0.004 + 0.001 + 0.003 = 0.008.$$

So, in crossovers with a binary response and a negative result, carryover effect should be tested by comparing the two periods with the less effective treatment modalities. If a significant test is demonstrated, we 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 is appropriate to disregard the data of the underneath 2nd period in this particular situation (although the 2nd period might still provide interesting information).

	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	$p_C = 0.2$

7 Discussion

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings, may all be adverse effects of an intervention on an outcome. An adverse effect will be called dependent, if it is significantly related not only to intervention, but also the outcome. In this chapter examples have been given, where carryover effect from previous treatments is an adverse effect of the effect of binary treatment modalities in crossover studies on the outcome. It is concluded, that, 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 is 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).

First, we have shown in the current chapter, that the adverse effect of carryover effect from the first into the second period of treatment in the control group can be easily tested for statistical significance. Second, we reasoned that, with such a significant carryover effect overall treatment assessment makes no sense, but we also gave solid arguments, that an unpaired analysis of the first periods of treatment is a worthwhile alternative for the purpose of assessment.

8 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 13
Confoundings and Interactions Assessed as Dependent Adverse Effects

Abstract With confounding a subgroup performs better for all treatments in a trial. With interaction a subgroup performs better for one treatment. Both confounding and interaction are adverse effects that may obscure the overall treatment efficacy of a trial, and they are adverse effects of the dependent type, i.e., they must be significantly related not only to the intervention but also to the outcome. In this chapter examples are given, as well as methods for assessment.

Also methods for assessing the presence of dependent adverse effects in *multi-* instead of *mono-* exponential mathematical models are discussed. The multiplication of such models as required for interaction/confounding assessments, is mathematically too complex for practical use, and they can, therefore, not be adequately used for assessing their presence. Alternatives, are, however, available.

Keywords Confounding · Subgroup · Interaction · Adverse effect · Obscured treatment efficacy · Adverse effects of the dependent type · Multi-exponential mathematical models

1 Introduction

Adverse effects are undesired effects with a medical intervention. Traditionally, they are supposed not to affect the main outcome of a study. Treatments produce, in addition to a main outcome effects, adverse effects. These adverse effects do or do not affect the main outcome. If they do, we'll call them dependent, if not, we'll call them independent. At current times multidimensional methodologies have brought to light novel types of adverse effects, that work directly on the main outcome of a study. Such adverse effects include not only confoundings and interactions, but also ancillary causal factors, categorical predictors, carryover effects (already reviewed in the Chaps. 10, 11 and 12). Why is it advantageous to interpret such divergent factors as adverse effects. This is because their mechanisms of action is pretty much similar, and so is their effect on the outcome of a study. This chapter will particularly address confoundings and interactions. With confounding a subgroup performs better for all

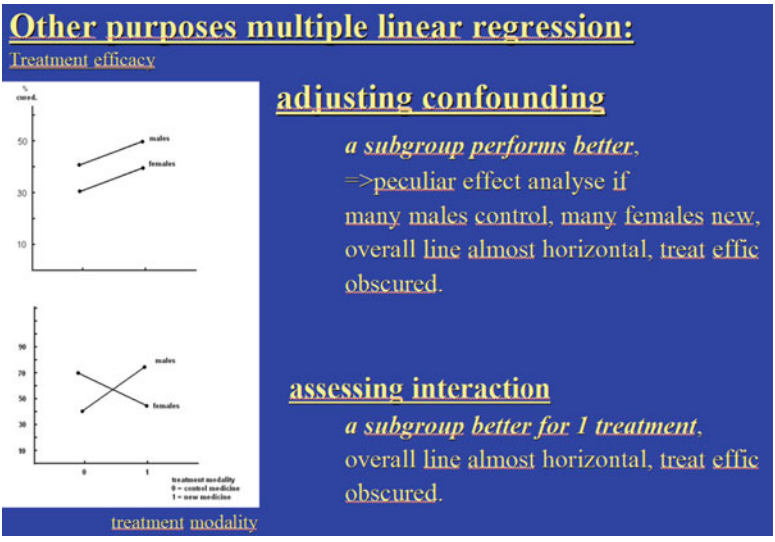
T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_13

treatments in a trial. With interaction a subgroup performs better for one treatment. This is explained in the underneath graph (effic = efficacy, treat = treatment). Both confounding and interaction are adverse effects that may obscure the overall treatment efficacy of a trial, and they are adverse effects of the dependent type, i.e., they must be significantly related not only to the intervention but also to the outcome. In this chapter examples will be given, as well as methods for assessment.

2 Difference Between Confounding and Interaction

The underneath figure, upper graph, shows an example of a parallel-group study of a control medicine and a new medicine, where confounding of genders is supposed to occur. The male subgroup performs better than the female does. This will have a peculiar effect on the efficacy analysis of the overall data, if we have many males in the control group and many females in new treatment group. Many males on control and females on new treatment means that the overall treatment modalities line will get almost horizontal, and the treatment efficacy will, thus, get obscured.

With interaction things are different. The underneath figure lower graph is an example. The males perform better on the new treatment, but the females do not so. Although the mechanism is different, again, the overall treatment modalities line will get almost horizontal, and, again, the overall treatment efficacy will get obscured. Both confoundings and interactions can assessed as dependent adverse effects.



3 Confounder as a Dependent Adverse Effect, Example

In a parallel-group study of the effect of two beta-blockers on cardiac output the two treatments were not significantly different. However, in the subgroup without bradycardia the results were slightly better than they were in the one with bradycardia.

		beta blocker 1	beta blocker 2
cardiac output (mean liter/min and standard error)			
bradycardia	yes	5.7 (0.8)	4.3 (0.5)
	no	6.1 (0.8)	4.7 (0.5)

In order to assess the presence of a significant confounding, subclassification was performed.

Dif 1 = mean difference between the bradycardia-yes treatment groups
Var 1 = variance of Dif 1

The weighted mean difference
$= (\text{dif } 1 / \text{var } 1 + \text{dif } 2 / \text{var } 2) / (1/\text{var } 1 + 1/\text{var } 2)$
$= (6.1\text{-}4.7) / (0.8^2 + 0.5^2) + \ldots\ldots / (1/0.89 + 1/0.89)$

$= 3.14606 / 2.247$
$= 1.4$

The standard error of the weighted mean difference
$= \sqrt{}\,[1 / (1 / \text{var } 1 + 1 / \text{var } 2)]$
$= \sqrt{}\,[1 / (1 / 0.89 + 1 / 0.89)\,]$
$= 0.67$

The t-test is used.
$t = 1.4 / .67$
$= 2.089$ (with 20 degrees of freedom $p < 0.05$)

Thus a significant confounding of the bradycardia variable is in the data. Confounding has a peculiar effect of the analysis. If you have many bradycardia-yes patients with beta blocker 1 and many bradycardia-no patients with beta blocker 2, then the treatment difference will be almost gone, and the overall treatment efficacy will be obscured. The bradycardia variable will be a significant adverse effect, and we will call it a *dependent* adverse effect, because it does not occur independently of the main outcome "cardiac output". So much so that the differences between the treatment modalities may be entirely lost. What is the solution? Perform separate analyses of the two bradycardia groups, because an overall analysis has become pretty meaningless.

4 Interaction as a Dependent Adverse Effect

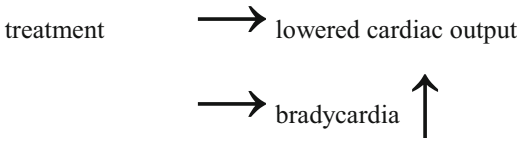
In a parallel-group study of calcium channel blocker versus beta-blocker for the treatment of paroxysmal atrial fibrillations (PAFs) the outcome was the numbers of PAFs, and the adverse effect was the presence of heart failure. Treatment may affect both PAFs and heart failure, while heart failure and PAFs were obviously not independent of one another. A t-test for interaction was performed.

	heart failure yes	heart failure no
mean numbers of PAFs (SD)		
calcium channel blocker	46.4 (3.24)	36.8(3.49)
beta blocker	30.2 (3.49)	37.8 (3.49)
Difference means (SE)	16.2 (1.51)	−10.0 (1.56)
Mean difference heart failure yes and no (SE) 17.2 (2.17)		
t = 17.2/2.17 = 8. . . .		
p<0.0001		

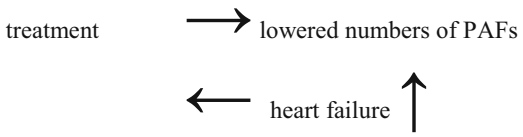
A significant difference in mean numbers of PAFs exists between heart failure yes and no patients. A significant interaction is between heart failure and treatment efficacy. The presence of heart failure is an adverse effect of treatment modalities on the main outcome "numbers of PAFs". we will call it a *dependent* adverse effect, because it does not occur independently of the main outcome "numbers of PAFs". So much so that the differences between the treatment modalities may be entirely lost. What is the solution? We recommend to perform here separate analyses of the two subgroups, the heart yes and the heart failure no subgroup, because an overall analysis has become pretty meaningless.

5 Causal and Inversed Causal Mechanisms

The relationship in the example of the above Sect. 13.3 between treatment and the adverse effect bradycardia may be causal as illustrated below.



However, an inversed causal mechanism between treatment and the adverse presence of the effect heart failure is also possible, as shown in the example of the above Sect. 13.4 between treatment and the adverse presence of the condition "heart failure" (as compared to the development of heart failure due to a treatment adverse effect (see below, and compare with the heart failure example in the Chap. 1).



6 Other Methods for Demonstrating Dependent Adverse Effects Due to Confounders and Interactions

Instead of subclassifications for confounders and t-tests for interactions, other assessments methods for the purpose are possible. For example analyses of variance and regressions can be used. These methods are more complex, and their mechanisms of action are less obvious. They will produce virtually the same results. In studies with more than a single confounder a regression analysis is possible with the treatment modality and the confounding variables as x-variables. With multiple confounders this method gets powerless, and propensity scores are a useful alternative. With propensity scores for every subgroup compute the chance of treatment 1 versus the chance of treatment 2, or, instead, the odds ratios of the two (used as relative chances). An example is given underneath (treat = treatment, DM = diabetes mellitus, ns = not significant).

	treat 1	treat 2	chance treat 1 / chance treat 2 or their Odds Ratios (OR)	
	n= 100	n = 100		p (vs OR 1)
1.Age>65	63 (%)	76	0.54 (63/37 / 76/24)	0.05
2.Age<65	37	24	1.85 (1/OR)	0.05
3.Dm	20	33	0.51	0.10
4.No DM	80	67	1.96	0.10
5.Smoker	50	80	0.25	0.10
6.No smoker	50	20	4.00	0.10
7.Hypertension	60	65	0.81	ns
8.No hypertension	40	35	1.23	ns
9.Cholesterol	75	78	0.85	ns
10.No cholesterol	25	22	1.18	ns
11.Renal failure	12	14	0.84	ns
12.No renal failure	88	86	1.31	ns

The ratios of the chance of treatment 1/chance of treatment 2 in various subgroups are computed using ORs, as surrogates for relative chances. With $p = 0.10$ as cut-off for statistical significance, conclude that age, diabetes mellitus, and smoker are dependent adverse effects. However, multiple testing is not addressed here. Propensity score adjustments or propensity score matching is a better alternative for the

purpose, and simultaneously addresses, and includes all dependent adverse effects due to confoundings in your analysis. With interactions, subgroups perform better on just a single treatment, and propensity scores are not adequate, but other methods are possible. We can assess, whether the differences-in-the-data due to interaction are large compared to those that are due to chance (residual). Analyses of variance and regressions can be conveniently used for the purpose. With large interaction effects, random effects analysis of variance and random effects regressions provide a more realistic and better sensitive result. More details will be covered in the Chap. 15.

With drug trials involving pharmacokinetic data special analytic models are required, that are unable to demonstrate dependent adverse effects. For example, pharmacokinetic/dynamic models, like the famous NONMEM models (nonlinear mixed effects models from the San Francisco University), can not be used for demonstrating dependent adverse effects. This is, because they apply *multi-* instead of *mono*-exponential mathematical models. The multiplications of such models, as required for interaction/confounding assessments, is mathematically too complex for practical use. Instead, multiple Cox models may be used, where concomitant predictors, like disease stage, and presence of special symptoms are included in the analysis. If such predictors are statistically significant, then the presence of dependent adverse effects can be demonstrated even so.

7 Discussion

Adverse effect are undesired effects with a medical intervention. Traditionally, they are supposed not to affect the main outcome of a study. Treatments produce, in addition to main outcome effects, adverse effects. These adverse effects do or do not affect the main outcome. If they do, we call them dependent, if not, we call them independent. At current times multidimensional methodologies have brought to light novel types of adverse effects, that work directly on the main outcome of a study. Such adverse effects include not only confoundings and interactions, but also carryover effects, random effects, unsafe patient characteristics, outlier data and more. Why is it advantageous to interpret such divergent factors as adverse effects. This is because their mechanism of action is pretty much similar, and so is their effect on the outcome of a study.

This chapter particularly addressed confoundings and interactions. With confounding a subgroup performs better for all treatments in a trial. With interaction a subgroup performs better for one treatment. Both of them are adverse effects that may obscure overall treatment efficacy. Both of them are adverse effects of the dependent type, i.e., they must be significantly related not only to the intervention but also to the outcome. In this chapter examples are given of confoundings and interactions, and methods for assessment.

Finally, we also discussed methods for assessing the presence of dependent adverse effects in *multi-* instead of *mono-* exponential mathematical models. The

multiplication of such models as required for interaction/confounding assessments, is mathematically too complex for practical use, and they can, therefore, not be adequately used for assessing their presence. Alternatives, are, however, available.

8 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 14
Subgroup Characteristics Assessed as Dependent Adverse Effects

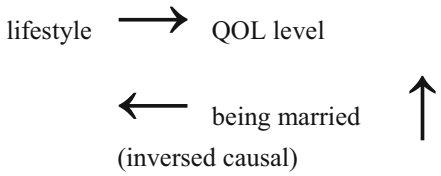
Abstract In this chapter examples are given where a subgroup characteristic like the effect of being married is a dependent adverse effect of the effect of, for example, life style on the outcome quality of life (QOL, qol).

Multinomial regression is traditional for identifying the main predictors of outcome categories, like levels of injury or quality of life (QOL, qol) categories. A better sensitive approach is various loglinear modelings. In the example given, logit loglinear, hierarchical loglinear of the orders 1st – 4th successively give increasing numbers of significant predictor interactions that can all be interpreted as dependent adverse effects.

Keywords Subgroup characteristic · Dependent adverse effect · Multinomial model · Logit loglinear model · Hierarchical loglinear models 1st to 4th order

1 Introduction

Causal relationships, pharmacological mechanisms, interactions, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings may all be dependent adverse effects of an intervention on an outcome Chaps. 1, 10, 11, 12 and 13). A dependent adverse effect must be significantly related not only to the intervention but also to the outcome. In this chapter examples will be given where a subgroup characteristic like the effect of being married is a dependent adverse effect of the effect of, for example, life style on the outcome quality of life (QOL, qol). The structural equation model is underneath.



T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_14

We should add, that being married is related to both lifestyle and QOL. The arrows indicate causal relationships. Being married might be a causal or inversed causal factor of lifestyle (see also Chap. 13, Sect. 13.5).

Multinomial regression is traditional for identifying the main predictors of outcome categories, like levels of injury or quality of life (QOL, qol) categories. An alternative and often better sensitive approach is loglinear modeling. It does not use continuous predictors on a case by case basis, but rather the weighted means of subgroups formed with predictors. This approach may allow for relevant additional conclusions from your data.

2 Multinomial and Logit Loglinear Models for Identifying Dependent Adverse Effects, an Example

In an example, the outcome was QOL levels, predictors were gender, married, lifestyle (0, 1).

Does logit loglinear modeling allow for relevant additional conclusions from your categorical data as compared to polytomous/multinomial regression? The data from the first 12 patients of the example are underneath.

qol	gender	married	lifestyle	age
2	1	0	0	55
2	1	1	1	32
1	1	1	0	27
3	0	1	0	77
1	1	1	0	34
1	1	0	1	35
2	1	1	1	57
2	1	1	1	57
1	0	0	0	35
2	1	1	0	42
3	0	1	0	30
1	0	1	1	34

age (years)
gender (0 = female)
married (0 = no)
lifestyle (0 = poor)
qol (quality of life levels, 1 = low, 3 = high)

The study included 445 patients. The characteristics are the predictor variables, the qol levels were the outcome. The entire data file is in extras.springer.com, and is entitled “loglinear”. We will first perform a traditional multinomial regression in order to test the linear relationship between the predictor levels and the chance (actually the odds, or to be precise logodds) of having one of three qol levels. Start by

opening SPSS, and entering the data file. For analysis the statistical model Multinomial Logistic Regression in the module Regression is required.

Command

Analyze....Regression....Multinomial Logistic Regression....Dependent: enter "qol".... Factor(s): enter "gender, married, lifestyle"....Covariate(s): enter "age"....click OK.

The underneath table shows the main results.

Parameter Estimates

qol[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
low	Intercept	28,027	2,539	121,826	1	,000			
	age	-,559	,047	143,158	1	,000	,572	,522	,626
	[gender=0]	,080	,508	,025	1	,875	1,083	,400	2,930
	[gender=1]	0[b]	.	.	0	.	.	.	.
	[married=0]	2,081	,541	14,784	1	,000	8,011	2,774	23,140
	[married=1]	0[b]	.	.	0	.	.	.	.
	[lifestyle=0]	-,801	,513	2,432	1	,119	,449	,164	1,228
	[lifestyle=1]	0[b]	.	.	0	.	.	.	.
medium	Intercept	20,133	2,329	74,743	1	,000			
	age	-,355	,040	79,904	1	,000	,701	,649	,758
	[gender=0]	,306	,372	,674	1	,412	1,358	,654	2,817
	[gender=1]	0[b]	.	.	0	.	.	.	.
	[married=0]	,612	,394	2,406	1	,121	1,843	,851	3,992
	[married=1]	0[b]	.	.	0	.	.	.	.
	[lifestyle=0]	-,014	,382	,001	1	,972	,987	,466	2,088
	[lifestyle=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: high.
b. This parameter is set to zero because it is redundant.

The following conclusions are appropriate.

1. The unmarried subjects have a greater chance of QOL level 1 than the married ones (the b-value is positive here).
2. The higher the age, the less chance of having the low QOL levels 1 and 2 (the b-values (regression coefficients) are negative here). If you wish, you may also report the odds ratios (Exp (B) values) here.

We will, subsequently, perform a logit loglinear analysis. For analysis the statistical model Logit in the module Loglinear in SPSS is used.

Command

Analyze.... Loglinear....Logit....Dependent: enter "qol"....Factor(s): enter "gender, married, lifestyle"....Cell Covariate(s): enter: "age"....Model: Terms in Model: enter: "gender, married, lifestyle, age"....click Continue....click

Options. . . .mark Estimates. . . .mark Adjusted residuals. . . .mark normal probabilities for adjusted residuals. . . .click Continue. . . .click OK.

The underneath table shows the results of the statistical tests of the data.

Parameter Estimates[c,d]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
[qol = 1]	5,332	8,845	,603	,547	-12,004	22,667
[qol = 2]	4,280	10,073	,425	,671	-15,463	24,022
[qol = 3]	0[b]	.	.	.	.	.
[qol = 1] * [gender = 0]	,389	,360	1,079	,280	-,317	1,095
[qol = 1] * [gender = 1]	0[b]	.	.	.	.	.
[qol = 2] * [gender = 0]	-,140	,265	-,528	,597	-,660	,380
[qol = 2] * [gender = 1]	0[b]	.	.	.	.	.
[qol = 3] * [gender = 0]	0[b]	.	.	.	.	.
[qol = 3] * [gender = 1]	0[b]	.	.	.	.	.
[qol = 1] * [married = 0]	1,132	,283	4,001	,000	,578	1,687
[qol = 1] * [married = 1]	0[b]	.	.	.	.	.
[qol = 2] * [married = 0]	-,078	,294	-,267	,790	-,655	,498
[qol = 2] * [married = 1]	0[b]	.	.	.	.	.
[qol = 3] * [married = 0]	0[b]	.	.	.	.	.
[qol = 3] * [married = 1]	0[b]	.	.	.	.	.
[qol = 1] * [lifestyle = 0]	-1,004	,311	-3,229	,001	-1,613	-,394
[qol = 1] * [lifestyle = 1]	0[b]	.	.	.	.	.
[qol = 2] * [lifestyle = 0]	,016	,271	,059	,953	-,515	,547
[qol = 2] * [lifestyle = 1]	0[b]	.	.	.	.	.
[qol = 3] * [lifestyle = 0]	0[b]	.	.	.	.	.
[qol = 3] * [lifestyle = 1]	0[b]	.	.	.	.	.
[qol = 1] * age	,116	,074	1,561	,119	-,030	,261
[qol = 2] * age	,114	,054	2,115	,034	,008	,219
[qol = 3] * age	,149	,138	1,075	,282	-,122	,419

a. Constants are not parameters under the multinomial assumption. Therefore, their standard errors are not calculated.
b. This parameter is set to zero because it is redundant.
c. Model: Multinomial Logit
d. Design: Constant + qol + qol * gender + qol * married + qol * lifestyle + qol * age

* = symbol of multiplication

The following conclusions are appropriate.

1. The unmarried subjects have a greater chance of QOL 1 (low QOL) than their married counterparts.
2. The inactive lifestyle subjects have a greater chance of QOL 1 (low QOL) than their adequate-lifestyle counterparts.
3. The higher the age the more chance of QOL 2 (medium level QOL), which is neither very good nor very bad, nut rather in-between (as you would expect).

We may conclude that the two procedures produce similar results, but the latter method provides some additional information about the lifestyle. We should note that multinomial regression is adequate for identifying the main predictors of outcome categories, like levels of injury or quality of life. An alternative approach is logit loglinear modeling. The latter method does not use continuous predictors on a case by case basis, but rather the weighted means of subgroups formed with the help of the discrete predictors. This approach allowed for relevant additional conclusions in the example given.

More background, theoretical and mathematical information of polytomous/multinomial regression is given in the Chapter 44 of SPSS for starters and 2nd levelers, Springer Heidelberg Germany, 2016, from the same authors. More information of loglinear modeling is in the Chapters 24 and 52 from the same edition.

3 Hierarchical Loglinear Interaction Models for Identifying Dependent Adverse Effects

Pearson chi-square test can anwer questions like: is the risk of falling out of bed different between the departments of surgery and internal medicine. The analysis is very limited, because the interaction between two variables is assessed only. However, we may also be interested in the effect of the two variables separately.

Also, higher order contingency tables do exist. E.g, we may want to know, whether variables like ageclass, gender, and other patient characteristics interact with the former two variables. Pearson is unable to assess higher order contingency tables.

Hiloglinear modeling enables to assess both main variable effects, and higher order (=multidimensional) contingency tables. For SPSS hiloglinear modeling SPSS statistical software does not provide a menu, but syntax is pretty easy, and syntax commands are given below. We should add, that hiloglinear modeling is the basis of a very new and broad field of data analysis, concerned with the associations between multidimensional categorical inputs. Can hierarchical loglinear modeling test all of the variable effects in multidimensional contingency tables? The above example was applied once more.

In 445 patients the effect of lifestyle (0 inactive, 1 active) on quality of life (qol) (0 low, 1 medium, 2 high) was studied. The marital status was considered to also affect the qol.

qol	age	gender	married	lifestyle outcome
2	55	1	0	0
2	32	1	1	1
1	27	1	1	0
3	77	0	1	0
1	34	1	1	0
1	35	1	0	1
2	57	1	1	1
2	57	1	1	1
1	35	0	0	0
2	42	1	1	0

The entire data file is in extras.springer.com, and is entitled "loglinear". Start by opening the data file in SPSS.

Analysis: First and Second Order Hierarchical Loglinear Modeling
Menu commands are not available. The syntax commands, however, are underneath.

Command
click File. . . .click New. . . .click Syntax. . . .Syntax Editor. . . .enter: hiloglinear qol (1,3) lifestyle (0,1)/criteria = delta (0)/design = qol*lifestyle/print = estim. . . .click Run. . . .click All.

The output sheets are underneath.

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	5	35,542	,000	35,391	,000	0
	2	2	24,035	,000	23,835	,000	2
K-way Effects[b]	1	3	11,507	,009	11,556	,009	0
	2	2	24,035	,000	23,835	,000	0

a. Tests that k-way and higher order effects are zero.
b. Tests that k-way effects are zero.

* = symbol of multiplication

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
qol*lifestyle	1	-,338	,074	-4,580	,000	-,483	-,193
	2	,246	,067	3,651	,000	,114	,378
qol	1	-,206	,074	-2,789	,005	-,351	-,061
	2	,149	,067	2,208	,027	,017	,281
lifestyle	1	,040	,049	,817	,414	-,057	,137

* = symbol of multiplication

The above tables in the output sheets show the most important results of the loglinear analysis.

1. There is a significant interaction "qol times lifestyle" at $p = 0.0001$, meaning that the qol levels in the inactive lifestyle group is different from those of the active lifestyle group.
2. There is also a significant qol effect at $p = 0.005$, meaning that medium and high qol is observed significantly more often than low qol.
3. There is no significant lifestyle effect, meaning that inactive and active lifestyles are equally distributed in the data.

Analysis: Third Order Hierarchical Loglinear Modeling
Command:

click File....click New....click Syntax....Syntax Editor....enter: hiloglinear qol (1,3) lifestyle (0,1) married(0,1)/criteria = delta (0)/design = qol*lifestyle*married/ print = estim....click Run....click All.

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	11	120,711	,000	118,676	,000	0
	2	7	68,839	,000	74,520	,000	2
	3	2	15,947	,000	15,429	,000	3
K-way Effects[b]	1	4	51,872	,000	44,156	,000	0
	2	5	52,892	,000	59,091	,000	0
	3	2	15,947	,000	15,429	,000	0

a. Tests that k-way and higher order effects are zero.
b. Tests that k-way effects are zero.

* = symbol of multiplication

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
qol*lifestyle*married	1	-,124	,079	-1,580	,114	-,278	,030
	2	,301	,079	3,826	,000	,147	,456
qol*lifestyle	1	-,337	,079	-4,291	,000	-,491	-,183
	2	,360	,079	4,573	,000	,206	,514
qol*married	1	,386	,079	4,908	,000	,232	,540
	2	-,164	,079	-2,081	,037	-,318	-,010
lifestyle*married	1	-,038	,056	-,688	,492	-,147	,071
qol	1	-,110	,079	-1,399	,162	-,264	,044
	2	,110	,079	1,398	,162	-,044	,264
lifestyle	1	,047	,056	,841	,401	-,062	,156
married	1	-,340	,056	-6,112	,000	-,449	-,231

* = symbol of multiplication

The above tables give the main results, and show that the analysis allows for some wonderful conclusions.

1. In the married subjects the combined effect of qol and lifestyle is different at p = 0.0001.
2. In the active lifestyle subjects QOL scores are significantly different from those of the inactive lifestyle subjects at p = 0,0001.
3. In the married subjects the QOL scores are significantly different from those of the unmarried ones at p = 0.037.
4. In the married subjects the lifestyle is not different from that of the unmarried subjects (p = 0.492).
5. The qol scores don't have significantly different counts (p = 0.162).
6. Lifestyles don't have significantly different counts (p = 0.401).
7. The married status is significantly more frequent than the unmarried status (p = 0.0001).

The many p-values need not necessarily be corrected for multiple testing, because of the hiearchical structure of the overall analysis. It start with testing first order models. If significant, then second order. If significant, then third order etc.

Analysis: Fourth Order Hierarchical Loglinear Modeling
Command:

click File....click New....click Syntax....Syntax Editor....enter: hiloglinear qol(1,3) lifestyle (0,1) married (0,1) gender (0,1)/criteria = delta (0)/ design = qol*lifestyle*married*gender/print = estim....click Run....click All.

K-Way and Higher-Order Effects

			Likelihood Ratio		Pearson		
	K	df	Chi-Square	Sig.	Chi-Square	Sig.	Number of Iterations
K-way and Higher Order Effects[a]	1	23	133,344	,000	133,751	,000	0
	2	18	81,470	,000	90,991	,000	2
	3	9	25,896	,002	25,570	,002	3
	4	2	,042	,979	,042	,979	3
K-way Effects[b]	1	5	51,874	,000	42,760	,000	0
	2	9	55,573	,000	65,421	,000	0
	3	7	25,855	,001	25,528	,001	0
	4	2	,042	,979	,042	,979	0

a. Tests that k-way and higher order effects are zero.
b. Tests that k-way effects are zero.

* = symbol of multiplication

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
qol*lifestyle*married*gender	1	-,006	,080	-,074	,941	-,163	,151
	2	-,010	,080	-,127	,899	-,166	,146
qol*lifestyle*married	1	-,121	,080	-1,512	,130	-,278	,036
	2	,297	,080	3,726	,000	,141	,453
qol*lifestyle*gender	1	-,096	,080	-1,202	,229	-,254	,061
	2	,086	,080	1,079	,281	-,070	,242
qol*married*gender	1	,071	,080	,887	,375	-,086	,228
	2	-,143	,080	-1,800	,072	-,300	,013
lifestyle*married*gender	1	-,065	,056	-1,157	,247	-,176	,045
qol*lifestyle	1	-,341	,080	-4,251	,000	-,498	-,184
	2	,355	,080	4,455	,000	,199	,511
qol*married	1	,382	,080	4,769	,000	,225	,540
	2	-,162	,080	-2,031	,042	-,318	-,006
lifestyle*married	1	-,035	,056	-,623	,533	-,146	,075
qol*gender	1	-,045	,080	-,565	,572	-,203	,112
	2	,018	,080	,223	,823	-,138	,174
lifestyle*gender	1	-,086	,056	-1,531	,126	-,197	,024
married*gender	1	-,007	,056	-,123	,902	-,118	,104
qol	1	-,119	,080	-1,488	,137	-,276	,038
	2	,111	,080	1,390	,164	-,045	,267
lifestyle	1	,041	,056	,720	,472	-,070	,151
married	1	-,345	,056	-6,106	,000	-,455	-,234
gender	1	-,034	,056	-,609	,543	-,145	,076

* = symbol of multiplication

The above tables show, that the results of the 4th order model are very much similar to that of the 3rd order model, and that the interaction gender*lifestyle*married*qol was not statistically significant. And, so, we can conclude here.

1. In the separate genders the combined effects of lifestyle, married status and quality of life were not significantly different.
2. In the married subjects the combined effect of qol and lifestyle is different at $p = 0.0001$.

3. In the active lifestyle subjects qol scores are significantly different from those of the inactive lifestyle at $p = 0.0001$.
4. The difference in married status is significant a $p = 0.0001$.
5. The qol scores don't have significantly different counts ($p = 0.164$).

The many p-values in the above analyses need not necessarily be corrected for multiple testing, because of its hierarchical structure. It start with testing first order models. If significant, then second order. If significant, then third order etc.

4 Discussion

Multinomial regression is a traditional method for for identifying the main predictors of outcome categories, like levels of injury or quality of life (QOL) categories. An alternative and often better sensitive approach is loglinear modeling. It does not use continuous predictors on a case by case basis, but rather the weighted means of subgroups formed with the help of predictors. This approach may allow for relevant additional conclusions from your data. In the current chapter it is shown, that, unlike multinomial regression, logit loglinear regression was able to demonstrate that lifestyle level zero was a very significant predictor of quality of life categories. Also, unlike multinomial regression, hierarchical loglinear regression was able to demonstrate that in the 1st and 2nd order models lifestyle was a very significant predictor of QOL, in the 3rd order model "lifestyle x being married" was so, and in the 4th order model "being married" was equally so. Several subgroups characteristics in a study of lifestyle on QOL were significant dependent adverse effects. The effect of being married was interpreted as inversed causal.

5 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 15
Random Effects Assessed as Dependent Adverse Effects

Abstract In this chapter examples are given, where a dependent adverse effect is random. In such studies, for example, a random "treatment by study subset" effect may be a dependent adverse effect of the treatment modalities on the outcome.

In clinical trials it is common to assume a fixed effects research model. This means that the patients selected for a specific treatment are 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 unexpected property.

Random effects models are a very interesting class of models, but even a partial understanding of it is fairly difficult to achieve. In this chapter we will demonstrate that it is helpful for the purpose to assess those random factors in the form of adverse effects of the dependent type. Random factor analysis implies that the treatment effect is not tested against the residual effect but rather against a random effect.

Keywords Dependent adverse effect · Random effect · Random "treatment by study subset" effect · Fixed effects research model · Unexpected property in data · Random factor analysis

1 Introduction

Causal relationships, pharmacological mechanisms, interactions, subgroup mechanisms, carryover effects from previous treatments, pleiotropic drug mechanisms, categorical factors, confoundings may all be dependent adverse effects of an intervention on an outcome (Chaps. 1, 10, 11, 12, 13, 14, 15 and 16). Unlike independent adverse effects, dependent adverse effects must be significantly related not only to the intervention but also to the outcome. In this chapter examples will be given

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_15

where the adverse effect is random. In such studies, for example, a random "treatment by study subset" effect may be a dependent adverse effect of the treatment modalities on the outcome.

2 Random Effects Research Models, Another Example of a Dependent Adverse Effects

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 unexpected 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.

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. Random effects models are a very interesting class of models, but even a partial understanding of it is fairly difficult to achieve. In this chapter we will demonstrate that it is helpful for the purpose to assess those random factors in the form of adverse effects of the dependent type.

3 A Random Effect of "Treatment by Study Subset" Assessed as a Dependent Adverse Effect

In a clinical trial with two study subsets 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. The random effect model is adequate for that purpose. An example is underneath. A parallel group study of two treatments consists of two subsets of patients. They were observed for numbers of episodes of paroxysmal atrial tachycardias after drug treatment.

Treatment		Verapamil	Metoprolol	
Study Subset 1	males	52	28	
		48	35	
		43	34	
		50	32	
		43	34	
		44	27	
		46	31	
		46	27	
		43	29	
		49	25	
		464	302	766
Study Subset 2	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	

The differences in the above data can be partitioned. The computations are (treat = treatment):

$$\text{SS total} = 52^2 + 48^2 + \ldots..35^2 - \frac{(52 + 48 + + \ldots 35)^2}{40} = 1750.4$$

$$\text{SS treat by study subset} = \frac{464^2 + \ldots 378^2}{10} - \frac{(52 + 48 + + \ldots 35)^2}{40} = 1327.2$$

$$\text{SS residual} = \text{SS total} - \text{SS treat by study subset} = 423.2$$

$$\text{SS gender} = \frac{766^2 + 746^2}{20} - \frac{(52 + 48 + + \ldots 35)^2}{40} = 10.0$$

$$\text{SS treat} = \frac{832^2 + 680^2}{20} - \frac{(52 + 48 + + \ldots 35)^2}{40} = 577.6$$

$$\begin{aligned}\text{SS interaction} &= \text{SS treat by study subset} - \text{SS gender} - \text{SS treat} \\ &= 1327.2 - 10.0 - 577.6 = 739.6\end{aligned}$$

Fixed effects analysis of variance

Source	SS	df	MS	F	p-value
Gender	10.0	1			
Treatment	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
Gender	10.0	1			
Treatment	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

Overall metoprolol scores performs better than verapamil, but this is only true for the patients in subset-1. The subset numbers seems to be a separate and unexpected variable in this study. The data are entered in the SPSS Software program 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) subset number. The software program enables to treat the independent variables either as fixed or as random variable. In the above table 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 subsets, 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 subset 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 would be required. The hypothesis of no treatment effect cannot be rejected. Thus, a significant “treatment by study subset” effect 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 A Random Effect of Health Center as an Adverse Effect of the Dependent Type

Underneath a table is given of a random effect study accounting the effect of the health center as random adverse effect in a study assessing the treatment on the outcome.

Treatment	Verapamil	Metoprolol	Isosorbide mononitrate	Total
	number of anginal 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:

$$\text{SS total} = 4^2 + \ldots. + 13^2 - \frac{(342)^2}{36} = 245$$

$$\text{SS treatment by health center} = \frac{10^2 + \ldots. + 27^2}{2} - \frac{(342)^2}{36} = 224$$

$$\text{SS residual} = \text{SS total} - \text{SS treatment by health center} = 245 - 224 = 21$$

$$\text{SS treatment} = \frac{99^2 + 105^2 + 138^2}{12} - \frac{342^2}{36} = 73.5$$

$$\text{SS health center} = \frac{49 + \ldots\ldots + 65^2}{6} - \frac{342^2}{36} = 30.67$$

$$\text{SS interaction} = \text{SS treatment by center} - \text{SS rows} - \text{SS columns} = 224 - 30.67 - 73.5 = 119.8$$

Fixed effects analysis of variance

Source	SS	df	MS	F	p-value
Health Center	30.67	6-1=5			
Treatment	73.5	3-1=2	36.75	36.75 /1.17= 31.49	< 0.0001
Interaction	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
Health Center	30.67	6-1=5			
Treatment	73.5	3-1=2	36.75	36.75 /11.98 = 3.07	ns
Interaction	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 = degree 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 (above table). 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 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 versus 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 Discussion

A unexpected effect like that of studies with subsets of patients implied in a single overall study may be an adverse effect of a treatment on the overall outcome that should be taken into account. A random effect analysis is adequate for the purpose. 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 gives examples of studies qualifying for them.

1. 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.
2. 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.

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.

In such studies, for example, a random "treatment by study subset" effect or "treatment by health center" may be a dependent adverse effect of the treatment modalities on the outcome. A random effect analysis can demonstrate whether the random effect is statistically significant.

6 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Chapter 16
Outliers Assessed as Dependent Adverse Effects

Abstract In a well-designed treatment trial the only difference between a treatment group and control group is the treatment. This is of course theoretically so. In practice many differences do exist, and raise the risk of biases.

Graphs like data plots and regression lines are convenient for visualizing outliers in therapeutic data patterns. Outlier data are considered as dependent adverse effects of the predictor data on the outcome data.

They are, however, arbitrary, and, with large data files, both data pattern and outlier recognition require a more sophisticated approach. Also, the number of outliers, generally, tends to rise with the sample size. BIRCH is the abbreviation of "balanced iterative reducing and clustering using hierarchies", and is available in SPSS's module Classify, under "two-step cluster analysis".

The current chapter, using a simulated and a real data example, examines whether BIRCH clustering is able to detect previously unrecognized outlier data. Step by step analyses were performed for the convenience of investigators.

Keywords Well-designed treatment trial · Risk of biases · Visualizing outliers · Outlier data · Dependent adverse effects · Predictor data · Outcome data · Outlier recognition · BIRCH "Balanced Iterative Reducing and Clustering Using Hierarchies · SPSS · Two step cluster analysis · Step by step analyses

1 Introduction

In a well-designed treatment trial the only difference between a treatment group and control group is the treatment. This is of course theoretically so. In practice many differences do exist, and do raise the risk of biases. Recruiting a homogeneous sample of patients is hard to do, and a prior heterogeneity assessment of the patient characteristics is helpful. An outlier assessment is a possibility for the purpose. Outliers can be determined, and removed from the data. In this chapter two examples will be given.

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3_16

First, in patients with mental depression two classes may be identified. One with endogenous depression and one with reactive depression. Endogenous depression causes severe depression in the younger, reactive depression causes mild depression in the elderly. Age may thus predict depression scores, and indeed in a sample of depressive patients age seemed to predict depression score in a regression model, although weakly. Insomnia is another producer of depression, and may be an outlier in a predictive study of the effects of age on depression. Birch clustering is helpful to demonstrate the presence of outliers in data. In a data example both younger patients with severe depression, and elderly with mild depression were observed. In addition, however, 14% of the patients were classified as an outlier category consistent of younger patients with mild depression and older patients with severe depression. The outlier data were considered as a dependent adverse effect of the predictor age on the outcome depression scores. The data example as given supports the above expected mechanism of action.

Second, in iatrogenic hospital admissions, age was a significant predictor of number of concomitant medications in a categorical model. Birch multidimensional clustering was able to identify not only clusters of young patients with few co-medications and older patients with many co-medications, but also a large outlier cluster of patients of all ages and "exceptionally-high-numbers-of-co-medications". This supports, that the cluster of patients at all ages and with very many co-medications is an outlier to be interpreted as a dependent adverse effect.

2 Birch Outlier Assessment

Graphs like data plots and regression lines are convenient for visualizing outliers in therapeutic data patterns, and have been successfully used for that purpose for centuries. They are, however, arbitrary, and, with large data files, both data pattern and outlier recognition require a more sophisticated approach. Also, the number of outliers, generally, tends to rise linearly with the sample size. BIRCH is the abbreviation of "balanced iterative reducing and clustering using hierarchies", and is available in SPSS's module Classify, under "two-step cluster analysis". It is an unsupervised data mining methodology suitable for very large datasets, but can also be applied for small data. It is, currently, mainly used by econo- and sociometrists, and, like other machine learning methods, little used in therapeutic research. This is, probably, due to the traditional belief of clinicians in clinical trials where outliers are assumed to be equally balanced by the randomization process and are not further taken into account. In contrast, modern computer data files often involve large uncontrolled data files, and arbitrary methods like scatter plots do not adequately detect outliers in the data.

The current chapter, using a simulated and a real data example, examines whether BIRCH clustering is able to detect previously unrecognized outlier data. Step by step analyses were performed for the convenience of investigators. This chapter was also written as a hand-hold presentation accessible to clinicians and a must read publication for those new to the method.

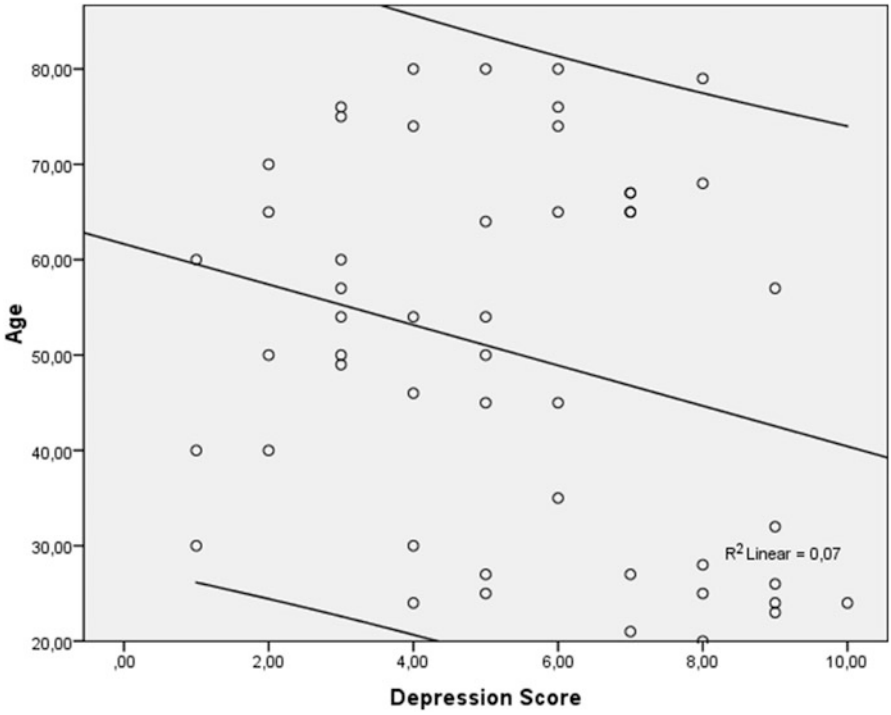
3 Example One

The underneath table shows a study of 50 mentally depressed patients. Age and depression severity scores (1 for mild and 10 for severest depression) are given in the first and second column. The cluster membership computed by two step BIRCH clustering is in column 3: two clusters were identified (indicated with 1 and 2) and one outlier cluster (indicated with -1).

Age	Depression score	Cluster membership
20,00	8,00	2
21,00	7,00	2
23,00	9,00	2
24,00	10,00	2
25,00	8,00	2
26,00	9,00	2
27,00	7,00	2
28,00	8,00	2
24,00	9,00	2
32,00	9,00	2
30,00	1,00	-1
40,00	2,00	-1
50,00	3,00	1
60,00	1,00	-1
70,00	2,00	1
76,00	3,00	1
65,00	2,00	1
54,00	3,00	1
54,00	4,00	1
49,00	3,00	1
30,00	4,00	2
25,00	5,00	2
24,00	4,00	2
27,00	5,00	2
35,00	6,00	2
45,00	5,00	1
45,00	6,00	2
67,00	7,00	1
80,00	6,00	1
80,00	5,00	1
40,00	1,00	-1
50,00	2,00	1
60,00	3,00	1
80,00	4,00	1
50,00	5,00	1
76,00	6,00	1
65,00	7,00	1
79,00	8,00	-1
57,00	3,00	1
46,00	4,00	1
54,00	5,00	1
74,00	6,00	1
65,00	7,00	1
57,00	9,00	-1
68,00	8,00	-1
67,00	7,00	1
65,00	6,00	1
64,00	5,00	1
74,00	4,00	1
75,00	3,00	1

Age and depression severity scores (1 for mild and 10 for severest depression) are given in the first and second column. Linear regression between the two variables gave some evidence for a weak negative correlation between the two with $p = 0.063$.

This would be compatible with the concept that younger are more at risk of high severity due to true depression, the older are so of low severity due to reactive depression. However, in case-reviews outlier forms of depression like insomnia groups have been noted, but no hints of such is given in the regression model. Even the 90% confidence intervals produced no more than a single case very close to the intervals boundary but otherwise no hint of outliers (figure below). An outlier analysis using two step BIRCH analysis was performed. SPSS statistical software was used for analysis.



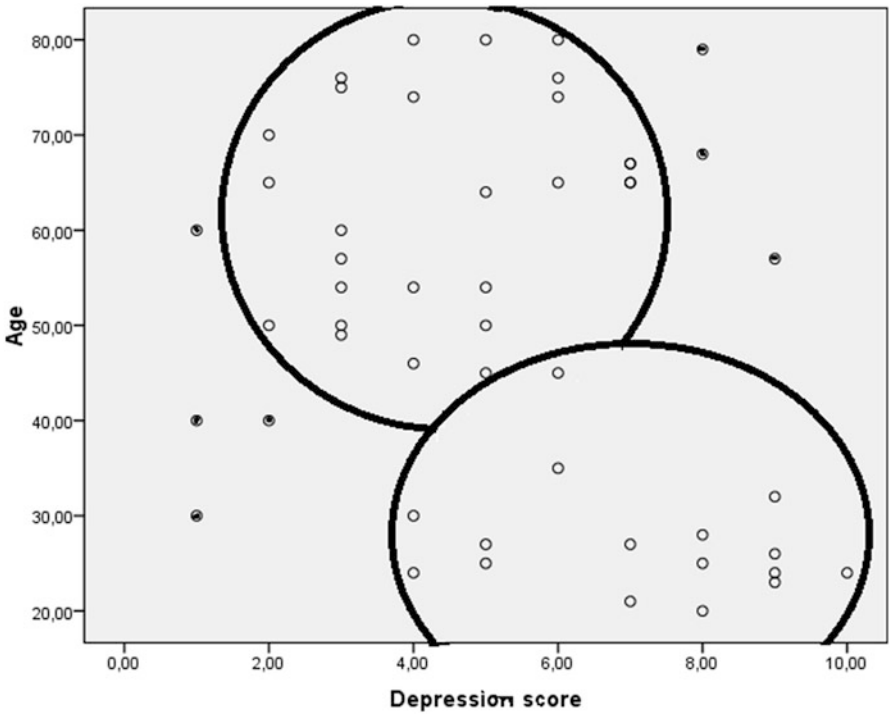
The SPSS data file is in extras.springer.com, and is entitled "birch1". Start by opening the data file in your computer mounted with SPSS statistical software.

Command Analyze. . . .Classify. . . .Two Step Cluster Analysis . . . .Continuous Variables: enter age and depression score. . . .Distance Measure: mark Euclidean. . . .Clustering Criterion: mark Schwarz's Bayesian Criterion. . . .click Options: mark Use noise handling . . . .percentage: enter 25. . . .Assumed Standardized: enter age and depression score. . . .click Continue. . . .click Output: Working Data File: mark Create cluster membership variable. . . .click Continue. . . . click OK.

When returning to the data file, it now shows the cluster membership of each case 1–50 (third column). Two clusters have been identified (indicated by 1 and 2) and one outlier cluster (indicated by −1). We will use SPSS again to draw a dotter graph of these results.

Command Analyze. . . .Graphs. . . .Legacy Dialogs: click Simple Scatter . . . . Define. . . .Y-axis: enter Age. . . .X-axis: enter Depression score. . . .OK.

The underneath figure shows two clusters with oval and, because of the similarly sized scales, even approximately round patterns. They are also approximately similar in size but this needs not to be so. Also, 7 outlier data are shown. The results do very well match the patterns as clinically expected: two populations, one with younger and severely patients with true depression and one with older and milder depressed patients with only a reactive depression. The outliers consist of 7 patients of all ages not fitting in the formed clusters. They may suffer from insomnia or other rare forms of the depression syndrome.



Thus, outlier detection using two step cluster analysis in SPSS identified two cluster and one outlier data set is. The lower cluster was compatible with younger

patients suffering from true depression, the upper cluster with older patients suffering from reactive depression. The outliers on the left and on the right side were 4 younger patients with low depression scores, and 3 older patients with high depression scores, and did not fit in the clusters formerly established.

4 Example Two

In a 2000 patient study of hospital admissions 576 possibly iatrogenic were identified by a team of specialists. The SPSS data file is in extras.springer.com and is entitled "birch2". The number of concomitant medications (co-medications) was not a significant predictor of hospital admission in the logistic regression of the data, but when transformed into a categorical factor it was. In order to find an explanation for this finding, a BIRCH two step cluster analysis of these data was performed in SPSS.

Open the data file in your computer mounted with SPSS statistical software.

Command Analyze. . . .Classify. . . .Two Step Cluster Analysis . . . .Continuous Variables: enter age and co-medications. . . .Distance Measure: mark Euclidean. . . . Clustering Criterion: mark Schwarz's Bayesian Criterion. . . .click Options: mark Use noise handling . . . .percentage: enter 25. . . .Assumed Standardized: enter age and co-medications. . . .click Continue. . . .click Plot: mark Cluster pie chart. . . .click Continue. . . .click Output: Statistics. . . .mark Descriptives by cluster. . . .mark Cluster frequencies. . . .mark Information Criterion . . . .Working Data File: mark Create cluster membership variable. . . .click Continue. . . .click OK.

The underneath table shows that 15 different cluster models have been assessed by the two-step BIRCH procedure (including 1–15 clusters). The precision of the different models, as estimated by the overall uncertainties measured by Schwarz's Bayesian Criterion (BIC) is given. With the 3 or 4 cluster models the smallest BIC was observed, and, thus, the most precise model.

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	293,899			
2	277,319	-16,580	1,000	1,513
3	185,362	-91,957	5,546	1,463
4	178,291	-7,071	,426	1,007
5	197,946	19,654	-1,185	1,141
6	216,403	18,457	-1,113	1,159
7	236,467	20,064	-1,210	1,099
8	251,072	14,606	-,881	1,629
9	272,582	21,509	-1,297	1,125
10	291,641	19,059	-1,150	1,015
11	301,090	9,449	-,570	1,000
12	308,019	6,929	-,418	1,058
13	321,943	13,924	-,840	1,197
14	339,382	17,439	-1,052	1,074
15	361,262	21,880	-1,320	1,225

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

The table below also in the output sheets, gives description information of the 4 cluster model selected from the 15 models from the above table.

Centroids

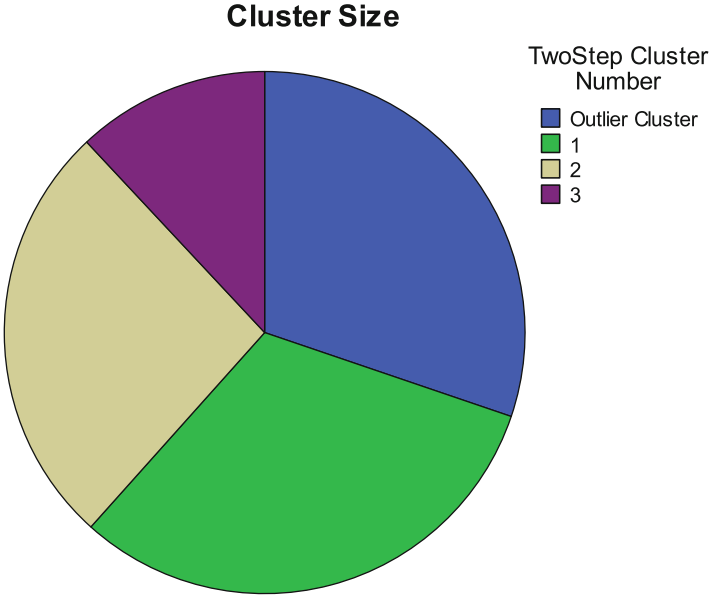
		age		comed	
		Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	1928,9227	6,50936	2,5028	,50138
	2	1933,7171	6,01699	,6250	,48572
	3	1956,8551	6,16984	1,0725	,64895
	Outlier (-1)	1939,7644	20,15623	2,4138	1,75395
	Combined	1936,8090	14,91570	1,8090	1,34681

In the table below are frequency information of the 4 cluster model selected from the 15 models.

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	181	31,4%	9,1%
	2	152	26,4%	7,6%
	3	69	12,0%	3,5%
	Outlier (-1)	174	30,2%	8,7%
	Combined	576	100,0%	28,8%
Excluded Cases		1424		71,2%
Total		2000		100,0%

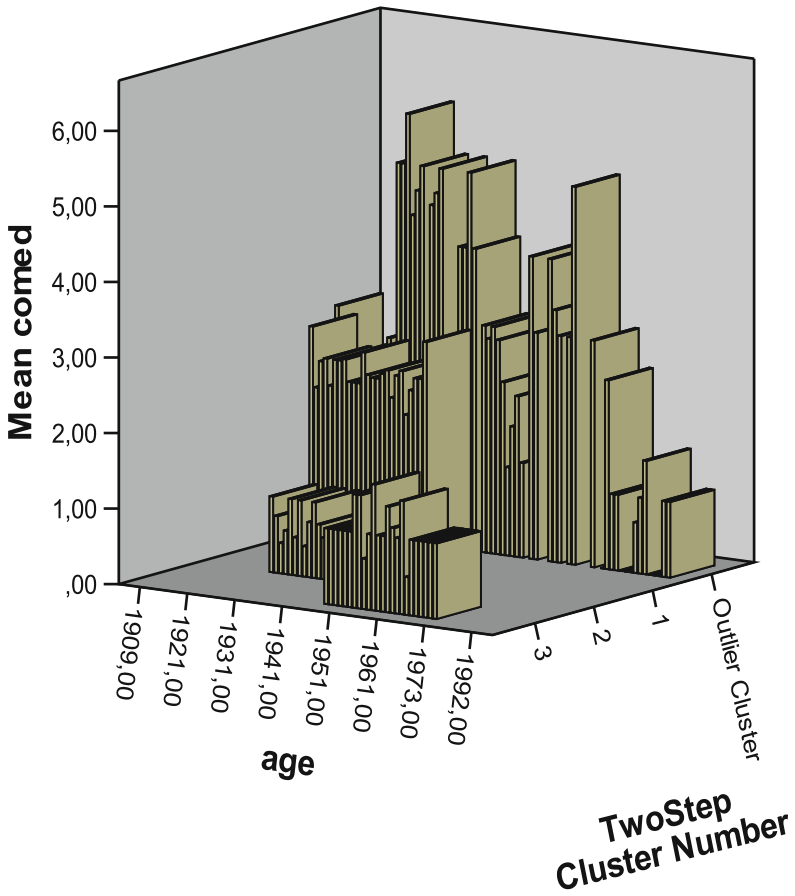
Thus, the above tables give the results of autoclustering of the two-step BIRCH procedure. It can be observed that 15 different models are assessed (including 1–15 clusters). Also is shown something about the precision of the different models, as estimated by the overall uncertainties (or standard errors) of the models (measured by Schwarz's Bayesian Criterion (BIC) = n ln (standard error)^2 + k ln n, where n = sample size, ln = natural logarithm, k = number of clusters). With the 3 or 4 cluster models the smallest BIC was observed, and, thus, the mostly precise model. The 3 or 4 cluster model, including an outlier cluster, would, therefore, be an adequate choice for further assessment of the data. Finally, description and frequency information of the 4 cluster model are given. The underneath figure in the output draws a pie chart of the size of the 3 clusters and the outlier cluster.



If we minimize the output pages, and return to the data file, we will observe, that SPSS has provided again the membership data. This file is too large to understand what is going on, and, therefore we will draw a three dimensional graph of this output.

Command Graphs. . . .Legacy Dialogs. . . .3 D Bar Charts. . . .X-axis represents: click Groups of cases. . . .Y- axis represents: click Groups of cases. . . .click Define. . . .Variable: enter co-medications. . . .Bars represent: enter mean of values. . . .X-Category axis: enter age. . . .Y-Category axis: enter two step cluster number variable. . . .click OK.

The figure below is shown in the output sheets. In front two clusters with younger patients and few co-medications are observed. In the third row is 1 cluster of elderly with considerably more co-medications. Then, at the back the patients are who do not fit in any of the clusters. They are of all ages, but their numbers of co-medications are generally very high. This finding is relevant, because it supports a deleterious effect of numbers of co-medications on the risk of iatrogenic admission.



The above three-dimensional bar chart is selected from the 4 cluster model. Over 100 bars indicate mean numbers of co-medications in age classes of 1 year. In the clusters 2 and 3 the patients are young and have few co-medications, in the cluster 1 the patients are old and have many co-medications, in the outlier cluster all ages are present and exceptionally high numbers of co-medications are frequently observed.

5 Discussion

There is no rigorous mathematical definition for outliers of a dataset, unlike there is for, for example, p-values, r-values etc. Why then worry about the outliers after all? This is, because they can lead not only to serious misinterpretations of the data, but also to catastrophic consequences once the data are used for making predictions, like serious and, sometimes, even fatal adverse events from drug treatments.

The current chapter shows that traditional methods like regression analysis is often unable to demonstrate outliers, while outlier detection using BIRCH two step clustering is more successful to that aim. We should add that this clustering method points to remote points in the data and flags them as potential outliers. It does not confirm any other prior expectation about the nature or pattern of the outliers.

The outliers, generally, involve both extremely high and extremely low values. The approach is, obviously, explorative, but, as shown in the examples, it can produce interesting findings, and theories, although waiting for confirmation. Other forms of cluster analysis include hierarchical, k-means and density-based clustering. Although they can produce multiple clusters, they do not explicitly allow for an outlier option. Nonetheless, investigators are, of course, free to make interpretations about outlier clusters from the patterns as presented.

This chapter only addresses two-dimensional data (one x and one y-variable), but, similarly to multiple regression, BIRCH analysis can be used for analyzing multi-dimensional data, although the computations will rapidly become even more laborious and computer memory may rapidly fall short. In the future this kind of research will be increasingly performed through a network of computer systems rather than a single computer system let alone standalone computers. Also, multidimensional outliers may be harder to interpret, because they are associated with multiple factors.

This chapter addresses only outlier-assessment in data without outcome variables. If outcome variables are available, other methods can be used, particularly, the identification of data beyond the confidence limits of the outcome variables. Also some special methods are possible, then. For example, looking for the data that are closer to expectation than compatible with random distributions, and investigating the final digits of the data values.

Outlier recognition and removal is an adequate method for identifying and adjusting the adverse effect of the predictors on the outcome of a study with heterogeneous subgroups. In a study where age is studies to predict numbers of co-medications, and adjusted and removed outlier cluster is helpful to adjust the adverse effect of age on the numbers of co-medications.

6 References

To readers requesting more background, theoretical and mathematical information of computations given, several textbooks complementary to the current production and written by the same authors are available:

Statistics applied to clinical studies 5th edition, 2012,
Machine learning in medicine a complete overview, 2015,
SPSS for starters and 2nd levelers 2nd edition, 2015,
Clinical data analysis on a pocket calculator 2nd edition, 2016,
Understanding clinical data analysis from published research, 2016,
Modern meta-analysis, 2017,
Regression analysis in clinical research, 2018,
Modern Bayesian statistics in clinical research, 2018.

All of them have been edited by Springer Heidelberg Germany.

Index

T. J. Cleophas, A. H. Zwinderman, *Analysis of Safety Data of Drug Trials*,
https://doi.org/10.1007/978-3-030-05804-3