Journal of
Clinical and Experimental Pharmacology

Gaus et al., Clin Exp Pharmacol 2015, 5:4
DOI: 10.4172/2161-1459.1000182

Research Article Open Access

Interpretation of Statistical Significance - Exploratory Versus Confirmative Testing in Clinical Trials, Epidemiological Studies, Meta-Analyses and Toxicological Screening (Using *Ginkgo biloba* as an Example)

Wilhelm Gaus, Benjamin Mayer and Rainer Muche

Institute for Epidemiology and Medical Biometry, University of Ulm, Germany

***Corresponding author:** Wilhelm Gaus, University of Ulm, Institute for Epidemiology and Medical Biometry, Schwabstrasse 13, 89075 Ulm, Germany, Tel : +49 731 500-26891; Fax +40 731 500-26902; E-mail: wilhelm.gaus@uni-ulm.de

Received date: June 18 2015; **Accepted date**: July 23 2015; **Published date**: July 27 2015

Abstract

The terms "significant" and "p-value" are important for biomedical researchers and readers of biomedical papers including pharmacologists. No other statistical result is misinterpreted as often as p-values. In this paper the issue of exploratory versus confirmative testing is discussed in general. A significant p-value sometimes leads to a precise hypothesis (exploratory testing), sometimes it is interpret as "statistical proof" (confirmative testing). A p-value may only be interpreted as confirmative, if (1) the hypothesis and the level of significance were established a priori and (2) an adjustment for multiple testing was carried out if more than one test was performed.

Screening programmes (e.g. the U.S. National Toxicology Programme on *Ginkgo biloba*) are typical for exploratory results. Controlled randomised trials include typically one confirmative test of the primary outcome variable and several exploratory tests of secondary outcome variables, as well as exploratory sub-group analyses. Some studies deliver p-values, which are more meaningful than merely exploratory, whilst other p-values appear to be more or less confirmative. Epidemiological studies and meta-analyses may lead to p-values, which are somewhat between exploratory and confirmative. We propose to consider exploratory and confirmative as a bipolar continuum. Nevertheless, authors of a study protocol are advised to design their study in a clearly exploratory or strictly confirmative manner. We furthermore recommend that each published significant p-value is explicitly denoted as exploratory or confirmative in addition to the appropriate descriptive results.

Keywords: Statistical test; Exploratory testing; Confirmative testing; p-value; Significance; *Ginkgo biloba*

Introduction

Estimation and hypothesis testing are cornerstones of statistics. Estimates and their confidence intervals deduce the size of e.g. risk factors and therapeutic effects and lead to "Statistics with Confidence" [1]. A confidence interval assesses how large the error of a point estimate maximally is, given a selected probability. But sometimes one has to come to a decision. Hypothesis testing is an often used possibility to handle inferential uncertainty rationally.

Today there is an intensive discussion on the usefulness of statistical testing and interpretation of p-values. This is also relevant for pharmacologists and toxicologists as it can be seen e.g. from the debate on *Ginkgo biloba* [2,3]. Our paper takes up this discussion and refers to a serious and constructive solution.

Two very basic principles have to be considered, when interpreting the result of a statistical test. The first rule is: "never interpret a p-value without the appropriate descriptive statistic." A test might be significant due to a large sample size, but the effect is too small to be relevant. In contrast, a remarkable effect is sometimes insignificant due to a small sample size or large variance. The second principle is to differentiate between exploratory and confirmative p-values. An exploratory interpretation of a significant p-value typically establishes a new hypothesis. By contrast, a confirmative interpretation of a significant p-value can be considered as "statistical proof" for a hypothesis established previously.

The misinterpretation of p-values is more and more recognised as a problem. Often significant p-values are interpreted as "statistical proof", even though the conditions for confirmative interpretation are not fulfilled. Some statisticians [4-6] recommend a more careful interpretation of p-values whilst others [7,8] advocate for the complete avoidance of significance testing. In accordance with [9,10] we believe that the latter approach is too restrictive. Instead, we suggest a more careful distinction between exploratory and confirmative significance testing.

An exploratory data analysis intends to identify all of the novel and interesting information contained in a data set. For this purpose, all possible comparisons shall be made. In addition to the formal term "exploratory data analysis", the terms "data exploration", "data snooping" and "data mining" are also used. Exploratory findings are necessary in order to point out new directions of research, to prepare for confirmative statistics, to give supportive information, to identify possible undesired effects of therapies and to find sub-populations, where a therapy is more or less efficacious.

Confirmative statistic is based on a detailed and specific hypothesis and aim to produce specific evidence. If the statistical test is significant, then the null-hypothesis is rejected, the alternative confirmed and a "statistical proof" achieved.

It is crucial to note that exploratory studies are not less important or easier to carry out than confirmative studies. Exploratory and

confirmative studies have different tasks and may be different types of investigation, but they do not differ in relevance, value and influence. Although exploratory and confirmative statistics are equally important, many researchers are content with being able to validate a supposed result by confirmative statistics.

The distinction between exploratory and confirmative statistics is not linked to the type of outcome variable (qualitative or quantitative), the number of groups (1, 2 or more groups), the design of the study (parallel groups or cross-over settings), or the statistical test applied (chi-square test, Wilcoxon test, analysis of variance, etc.).

The purpose of this paper is to sensitise users of statistics, e.g. pharmacologists, toxicologists, experimental researchers and clinicians, towards a distinction between exploratory and confirmative p-values. An inadequate interpretation of statistical findings can result in false benefits or risk assessments, which is of course especially serious in the health care sector.

Illustrative Example

In order to demonstrate the dramatic difference between an exploratory and confirmative interpretation of statistical test results, let us consider a fictitious conversation between a clinician and a statistician, which commonly occurs in practice.

A clinician, let us call him John Busy, has collected data from all the patients in his hospital with inguinal hernias over the last four years. He assigned the patients to four groups: 156 men ≥ 40 years, 121 men < 40 years, 48 women ≥ 40 years and 43 women < 40 years. After admission and before discharge he recorded the following seven variables for each patient: serum creatinine, bilirubin, gamma-GT, LDL, haemoglobin, haematocrit and CRP. Additionally, he recorded the length of stay. For each variable, he computed the mean and standard deviation for each of the four groups and the two time points. In total there are 60 means (7 variables × 4 groups × 2 points of time + length of stay for 4 groups) and the associated 60 standard deviations.

John Busy looked carefully at his results and compared them in many ways. He found three remarkable differences of means as follows: (1) the Gamma-GT between admission and discharge for men ≥ 40 years, (2) the length of stay between females < 40 years and males < 40 years, (3) the serum creatinine at discharge between women ≥ 40 years and women < 40 years.

John then approached the biometrical institute with his data and the aforementioned results. There, he asked the statistician Fred Fussy to check, whether these three observed differences were significant. Fred Fussy mentally computed a rough estimate t-test for each of the three remarkable differences of two means found by John Busy. He then told John: "had you not seen the data and the means, then all three differences would be significant." (Throughout this conversation, Fussy uses the term significant in a confirmative sense.) "But since you have already seen the data, the means and the standard deviations, these three differences are not significant." John Busy replied: "But I never manipulated the data nor the results." "I believe you," answered Fussy "however, because you have seen the descriptive statistical results before you consulted me, the results are not significant."

Dear reader, you may now have sympathy with John Busy and might think that Fred Fussy has lost his mind. Please read this paper in its entirety in order to also understand the statistician's arguments [11].

Methods

How many tests on pairwise comparisons are possible with a given data set?

Let us assume that a researcher has from an observational study a data set with v variables, observed for g groups at t time points. How many pairwise comparisons are possible?

Each of the g groups can be compared to each other group, resulting in g (g−1) comparisons. Since the comparison of group A versus group B provides the same information as the comparison of group B versus group A, only

½ g (g − 1) pairwise comparisons are possible in regards to the groups.

Each of the t time-points can be compared to the other respective time point. With thought-patterns, which are analogue to the groups, we can identify

½ t (t − 1) pairwise comparisons in regards to the time points.

Each of the v variables can be evaluated separately. The comparisons between the two groups and the comparisons between the two time points can be carried out for each variable. Adding up and resolving the components lead to a

total number of possible pairwise comparisons = ½ v g t (g + t − 2)

When applying this formula to the data set of John Busy with v=7 variables, g=4 groups and t=2 time points, this leads to

½ × 7× 4 × 2 × (4 + 2 − 2) = 112 possible pairwise comparisons.

Additionally, there is the length of stay for the four groups. This enables ½ × 4 × (4 − 1) = 6 pairwise comparisons in addition. Thus, John's data set allows 112 + 6 = 118 possible pairwise comparisons.

Today most data sets are much larger than John's regardless if they are from clinical studies, animal research, cell cultures or *in vitro* experiments. This increases the number of comparisons once again.

In order to simplify the problem, we shall only discuss pairwise comparisons. If other comparisons are considered as well, e.g. group A and B are pooled and then compared to group C, then the number of possible comparisons and therefore the number of possible tests would further increase. All possible comparisons are handled by closed test procedures including the global test, all intersection hypotheses and all pairwise comparisons [12]. Following this, seven comparisons between groups are considered for three groups, and 14 comparisons between groups are considered for four groups.

Expected number of false significant tests in an exploratory data analysis

The level of significance of a statistical test is the probability that the test gives a significant result even though there is no actual effect. Therefore, if a test is applied to data stemming from a random number generator, i.e. to data without any effect, then the probability to achieve a false significant result is the selected level of significance.

We assume that a data set is completely generated by a random number generator. Then the null hypothesis is actually valid for all possible comparisons. If n tests are possible within the data set and if for all these n tests the level of significance of α was established, then

we expect $n \times \alpha$ significant results. All of these significant results are false, because we applied all of these tests to random numbers.

An exploratory data analysis deals with many questions and problems. The data exploration will identify any interesting result. Thus – formally – all possible tests for the data set would be carried out. Nobody will of course actually compute tests for all comparisons possible in a data set. If it can be seen from the descriptive results that the difference is rather small and that it is unlikely that significance will be achieved, then researchers will save superfluous work and will not compute the appropriate test. It is not relevant, if a test is actually computed or if the computation of the test is omitted due to poor descriptive results. Hence, for the computation of the expected number of false significant tests, only the number of possible tests in the data set is relevant, not the number of actually computed tests.

Let us go back to the example of John Busy. In his data set, it was possible to carry out 118 tests in regards to the pairwise comparisons. It is assumed that the usual level of significance of $\alpha = 5\%$ was chosen for all tests. If the whole data set is actually without any effect, we expect $118 \times 0.05 = 5.9$ false significant results.

An exploratory significance generates a hypothesis

In experimental research, one has to strictly distinguish between two steps: hypothesis generation and hypothesis confirmation. It is very important that the data for the hypothesis confirmation is independent from the data used for the generation of the hypothesis.

However, Bretz et al. [13] and Stallard [14] published with the seamless design an exception from this rule. A phase II study on the efficacy of a therapy may be carried on as a phase III study, if the p-value is properly adjusted. In order to maintain clarity and simplicity in our manuscript, we only mention this once. At all other locations of this manuscript we use the classical approach that hypothesis generation and hypothesis confirmation have to be independent of each other.

A hypothesis should be a precise and detailed description of how reality could be. The creation of an accurate and promising hypothesis is an important step. A hypothesis can be gained by intuition or through theoretical considerations, but mostly it is generated by an exploratory analysis of data. By extracting a substantiated and detailed hypothesis from an exploratory study, half of the research work is already done! Now, an experiment to confirm the established hypothesis may be planned. Thus, exploratory research for the generation of hypotheses is equally important as the confirmation of hypotheses.

In principal, an exploratory analysis can only generate hypotheses, but it can never prove a hypothesis. If an exploratory significance has been obtained, then a hypothesis is generated. However, descriptive statistical results and other considerations should be used to decide, whether further research in order to confirm this hypothesis is worthwhile.

In the illustrating example, the statistician could not see that John Busy had established any hypotheses beforehand. Thus, he assumed an exploratory data analysis. If three tests are actually significant, but 5.9 false significant tests were expected in the case that there are no effects at all in the data set, then the achieved three significant results are less important. But Fred Fussy was imprecise when he said that there is no significance. There are indeed three significant results. But they are exploratory, i.e. John Busy generated three novel hypotheses through his study.

Confirmative testing needs a hypothesis and a level of significance both established a priori

"A priori" means that the hypothesis was established beforehand, so prior to the actual experience. Several years ago, Hanns Klinger (Düsseldorf, Germany) said half-joked that he would only accept a test result to be confirmative, if the hypothesis was deposited by a notary public before beginning the study. Today study protocols and their hypothesis are registered e.g. in *ClinicalTrials.gov*. In order to carry out confirmative testing for an a priori established hypothesis, new data is required, which was not available at the time of generating the hypothesis.

The level of significance has to also be established beforehand.

Furthermore, one needs to remember the saying that a precise answer (confirmative testing) is only possible for a precise question (hypothesis).

Number of chances and multiple testing

Confirmative testing has a third prerequisite. Strictly speaking, to facilitate a confirmative interpretation only one chance to identify a significant result is allowed, i.e. only one test is permitted. The more hypotheses that were established a priori, the more chances exist to achieve a significant result, and the more we approach exploratory statistics.

Today, however, there are several methods available to adjust for multiple testing [15,16], e.g. to split the level of significance to several tests according to Bonferroni, the Bonferroni-Holm procedure and the Hochberg procedure. Alternatively, the concepts of gatekeeping, a priori ordered hypotheses or the closed test procedure can be applied in order to control the family wise error rate. All of these approaches reduce the chances of a false significant result. The more hypotheses that are being tested – regardless of whether the test has actually been computed or whether it could be seen in the descriptive statistical results, making it unnecessary to carry out any computational work – the stronger the adjustment has to be in order to allow for a significance being interpreted as confirmative.

If all three prerequisites are fulfilled, (1) if a precise hypothesis and (2) the level of significance have been established independently from the data the hypothesis is tested and (3) if – in case of more than one test was scheduled – the p-values were adjusted for multiple testing, then – and only then – may a significant test result be considered as confirmative, as a "statistical proof" so to say.

All studies which do not fulfil one or more of the above mentioned three prerequisites are exploratory. Thus, there is a variety of explorative studies. Sometimes a significant p-value supports the main result of a study. But is this support exploratory or confirmative? If the supporting result was under investigation beforehand (and type I error rate was controlled for it) then it may be interpreted as confirmative. However, if the supporting result was unexpected at the onset, then it is exploratory. A guideline of the European Medical Agency (EMA) on "Points to consider on multiplicity issues in clinical trials" [17] gives more information on this topic.

Exploratory and confirmative as a bipolar approach

The results of a study are not always unequivocal exploratory or confirmative, although an exploratory or strictly confirmative test was intended in the study protocol. Many problems may occur during the conduction of a larger trial. After an evaluation, it is often discussed whether the achieved significance should be interpreted as exploratory or confirmative. A terminated study can be somewhat more meaningful than being solely exploratory. A study could have been planned as a confirmative one, but it encountered problems and finally, only a tendency could be given. Significance can be more or less exploratory or confirmative. Sometimes, an obtained significance is despite some efforts to achieve pure results somewhere in between exploratory and confirmative. Especially meta-analyses and epidemiological studies are not always indisputable exploratory or confirmative. For more detailed examples see section Interpretation of studies. However, when planning a survey or trial, it should be clearly defined whether it will lead to an exploratory or confirmative result.

We therefore believe that exploratory and confirmative are two poles with a continuum in between. The exploratory interpretation is one pole position, and the confirmative interpretation the other. Some performed studies are actually more or less exploratory, whilst other studies are more or less confirmative. Some studies are in between these poles. Of course, hybrid significance should be avoided as far as possible. Figure 1 provides results for 100 fictitious studies.

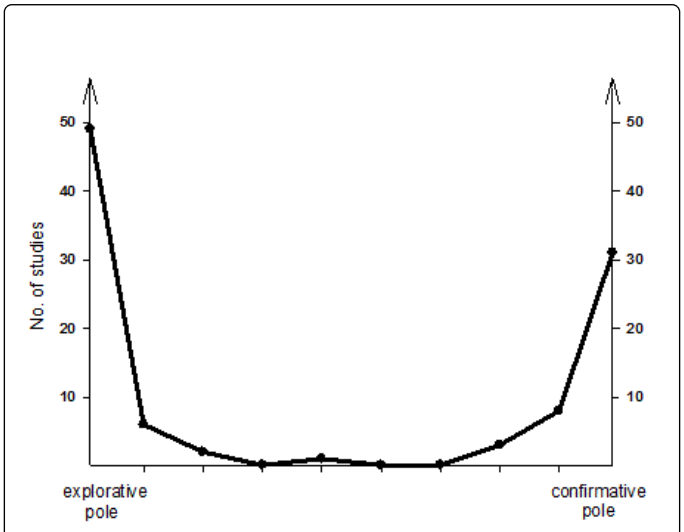

Figure 1: Many significant results are truly exploratory, and many significant results are strictly confirmative. But some significant results are more relevant than mere exploratory. Some other significant results are confirmative only in tendency, but not from a strict point of view. A few studies are in between. Typically, meta-analyses are only confirmative in tendency, whilst screening investigations are always close to the exploratory pole. The diagram shows the results for 100 fictitious studies.

Interpretation of studies

Controlled randomised clinical trials

The usual controlled randomised trial is confirmative and exploratory at the same time. Typically, one specific a priori established hypothesis defines the groups that are to be compared, such as type, dosage and duration of treatment, the point in time of the comparison, the primary outcome variable, as well as the level of significance. This hypothesis is then tested and the significance may be interpreted as confirmative. But the study could also investigate several secondary outcome variables, various sub-group analyses and different points in time. These tests are exploratory.

Studies with several a priori established hypotheses

The following situation is assumed: a researcher plans an experiment, defines his data set, and includes many or even all possible hypotheses with this data set in the study protocol. Additionally, he defines the level of significance for each hypothesis. Thus, all of the hypotheses are formally established a priori. At the evaluation stage, each of the hypotheses, which have been formulated in the protocol, is tested with one single test, if the appropriate descriptive statistic indicates a promising effect. No other tests are carried out. Thus, all of the prerequisites for confirmatory testing – a priori formulated hypothesis and level of significance, as well as one test only for each hypothesis – are fulfilled! Are the significant results exploratory or confirmative?

Let us compute the expected number of significant tests under the assumption that all of the data stems from a random number generator, i.e. the null-hypothesis is actually valid for all of the established hypotheses. If the number of observed significant results is around the number of expected false significant tests, then it is obvious that the significant p-values are exploratory. If the number of observed significant tests is much larger than the expected number of false significant tests, then it remains open, which of the observed significant tests are correctly significant. We would label this study as an exploratory investigation, due to the large number of chances to achieve significant results.

Let us now consider a slightly modified example. It is the same situation as above, but in addition, the study protocol contains an adjustment of the p-values using the Bonferroni-Holm procedure. The number of tests inserted for the Bonferroni-Holm procedure does not correspond to the number of tests actually computed, but to the number of all a priori established hypotheses. Are the significant results exploratory or confirmative in this case? We believe this to be more of a confirmative investigation. An adjustment for multiple testing is necessary, if different problems and hypotheses are tested in one study, in order to interpret the significant results as confirmative.

Very small exploratory p-values

Sometimes, a very small p-value occurs in exploratory evaluations. Is such an extremely small p-value already confirmative? The smaller an exploratory p-value is, of course, the better the chance for significance in the following confirmative study. But Bretz and Westfall [18], using a large simulation, show that in most cases, a small exploratory p-value cannot be repeated in the following confirmative study. Ring and Eskofier [19] achieved a similar result for the U.S. National Toxicity Program (NTP) [20].

In a typical exploratory situation, it is not known beforehand how many tests will be feasible. Therefore, an adjustment for multiple testing is not possible. More important, the chance, that a subsequent confirmatory study is successful, can be better predicted by using the effect size estimated by the exploratory study rather than by using its p-value. An estimation of the effect size is furthermore needed in order

to calculate the sample size of the subsequent confirmatory study, rather than the p-value.

Very small probabilities are rare, but do sometimes occur. Let us look at an example for a very small probability. In gambling, the probability to collect the jackpot is indeed very small, but sooner or later, someone will win it. Thus, even extremely small exploratory p-values are not confirmatively significant.

Epidemiological studies

Cohort studies and case-control-studies are common in epidemiology. A cohort study investigates one (or a few number of) exposure(s) and the many possible consequences of this exposure. A case-control-study investigates one problem and looks for many causes. Thus, cohort studies screen for events or failures, whilst case-control-studies screen for causes and reasons. However, a type II error is often worse in epidemiology compared to a type I error. We think that cohort studies and case-control-studies are somewhere in between the two pole positions of exploratory and confirmative. This could be the reason for many epidemiologists preferring confidence intervals instead of significance testing.

Meta-analyses

A meta-analysis combines the results of several studies on the same topic in order to come to a final decision. A typical indication is that in all studies the sample size was too small to deliver a significant result. But by pooling these studies, a significant result could be achieved.

Gaus et al. [21] ask whether meta-analyses are exploratory or confirmative. If the studies included in a meta-analysis are exploratory, then the results of the meta-analysis would also be exploratory. But usually, the studies included in a meta-analysis have a confirmative design. If all studies incorporated in a meta-analysis are confirmative and investigate the same patient population, the same treatment conditions, the same treatment duration, as well as the same primary outcome variable, then the meta-analysis may also be considered as being confirmative.

The protocol of a meta-analysis is typically written, when the data of the planned meta-analysis, i.e. the results of the single studies, are at least partly known. The meta-analyst is able to look at the results of the different outcome variables of the published studies already known to him. He then selects the outcome variable for his meta-analysis in order to enable the best chance for achieving a significant result. Thus, the design of the meta-analysis partly depends on the results of the studies to be meta-analysed. This is typical for exploratory statistics.

A lazy answer would play it safe and state that meta-analyses are always exploratory. But we believe the answer to be more complicated. We recommend the avoidance of general decisions, and to instead make individual decisions for each meta-analysis, if its primary result has to be considered as either exploratory or confirmative.

Toxicological screening programmes

The purpose of a toxicological screening programme is to identify interesting matters of all types in a considered field. Usually, when a toxicological screening programme is launched, there is some hope, but no precise idea in regards to the expected findings. Not all of the details of the investigational programme can be defined beforehand in a typical screening investigation. If a number of interesting topics are identified during a screening programme, further investigations are to be carried out in this respect. In contrast, if there are no interesting results, then further investigations will be omitted. Thus, the number of investigations – and more precisely the number of possible statistical tests – cannot be defined beforehand. The words "to screen" and "to explore" already have a similar meaning. Thus, toxicological screenings are typically exploratory.

Many toxicological experiments have the following structure: animals of a certain species (mostly rodents) are exposed to a test substance for some time. Often, the exposure is a specific diet. Most experiments have several groups. Each group receives a defined dose of the exposure. The group with dose zero is the control group. After the defined time of exposure, the animals are sacrificed and all of the organs are fully investigated, both macroscopically and microscopically. Typically, 10 to 100 tissues are investigated and for each tissue, 10 or more different findings are possible. In total, between 100 and up to several thousand single investigations are recorded. Thus, the experiment produces a rather large data set for exploratory tests.

Example: US National Toxicology Programme (NTP) on *Ginkgo biloba*

A typical toxicological screening programme is the U.S. National Toxicology Programme (NTP), e.g. the Technical Report (TR) 578 on *Ginkgo biloba* [20]. In this report, a 3-month study on rats, a 3-month study on mice, a 2-year study on rats, and a 2-year study on mice are being reported on. In the first mentioned study groups of 10 male and 10 female rats were administered 0, 62.5, 125, 250, 500, or 1 000 mg of *Ginkgo biloba* extract / kg body weight in corn oil by gavage, 5 days a week for 14 weeks. At the end of the experiment, the stomach, liver, bile duct, thyroid gland, kidney, nose, and other locations were investigated. Typical findings were hypertrophy, atrophy, hyperplasia, inflammation, hyperkeratosis, ulcers, pigmentation etc.

There is now a serious discussion [19,22] on the evidence produced by the NTP, especially in regards to TR 578 on *Ginkgo biloba* [20]. Obviously, NTP is a screening programme for cancer and other pathological findings. Any observed irregularity is of interest. Before the experiment, the idea was to investigate, whether there were any irregularities and if so, which types they would be. There are many possible statistical tests. Gaus [2] points out that the number of reported significant findings is in the same size order as the number of expected false significant tests. He therefore concludes that all significant findings in TR 578 [20] are exploratory and that they generate a hypothesis. Kissling, Haseman and Zeiger - who are more or less involved with the NTP – reject this idea passionately [3]. They claim that the NTP does not only generate hypotheses, but that it also proves at least some of them. As an example, they mention an extremely small p-value and argue that it is not only exploratory but confirmative as well, as it is so small. Unfortunately, this very small p-value is neither mentioned in [20], nor provides [3] a corresponding descriptive statistic, nor mention [3] any hypotheses established beforehand. This discussion may be an indication that the distinction between exploratory and confirmative testing is not yet widespread knowledge.

Recommendations

All authors reporting on experiments and studies, as well as all readers of statistical results, have to know whether the reported results are exploratory or confirmative. It is a serious discrepancy, if a result

generates a hypothesis or if an a priori established hypothesis is confirmed. Unfortunately, not all authors disclose to their readers, whether a reported p-value is exploratory or confirmative. Statisticians should pay more attention to the following two points: (1) each study protocol should clearly declare, which tests are exploratory and which tests are confirmative; (2) authors should state for each significant p-value, whether it is exploratory or confirmative in all reports and publications. We furthermore support the rule that every p-value should be accompanied by an appropriate descriptive statistic, although this is out of the scope for this paper.

Not all of the significant results are clearly exploratory or undoubtedly confirmative. Some are somewhere in between exploratory and confirmative. We propose to consider exploratory and confirmative as two pole positions with a continuum in between. This is especially helpful in regards to the assessment of meta-analyses and epidemiological studies. However, such hybrid significance testing should be avoided as far as possible.

Many high-level journals recommend to authors of clinical trials to report according to the CONSORT statement [23]. This essentially improves a reader's assessment of the validity of the design, the performance and the results of the studies. In "Table 1 CONSORT checklist of information to include when reporting a randomised trial" point 18 reads "Results of any other analyses performed, included subgroup analysis and adjusted analyses, distinguishing pre-specified from exploratory" We propose to add the differentiation between exploratory and confirmative significance in the CONSORT-statement more general and striking.

Acknowledgement

We thank the anonymous reviewer for her or his constructive and helpful comments.

References

1. Altman DG, Machin D, Bryant TN, Gardner MJ (2000) Statistics with confidence. BMJ books, 2nd edition.
2. Gaus W (2014) Which level of evidence does the US National Toxicology Program provide? Statistical considerations using the Technical Report 578 on Ginkgo biloba as an example. Toxicol.Lett. 229: 402-404.
3. Kissling GE, Haseman JK, Zeiger E (2014) Proper interpretation of chronic toxicity studies and their statistics: A critique of "Which level of evidence does the US National Toxicology Program provide? Statistical considerations using the Technical Report 578 on Ginkgo biloba as an example". Toxicol Lett. 237:161-164.
4. Nuzzo R (2014) Statistical errors. P-values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506: 150-152.
5. Victor A, Elsasser A, Hommel G, Blettner M (2010) Judging a plethora of p-values: how to contend with the problem of multiple testing--part 10 of a series on evaluation of scientific publications. Dtsch Arztebl Int 107: 50-56.
6. Leek JT, Peng RD (2015) Statistics: P values are just the tip of the iceberg. Nature 520: 612.
7. Hak T (2014) After statistics reform: Should we still teach significance testing? In: Makar K, de Sousa B, Gould R (Eds.) Sustainability in statistics education. Proceedings of the 9th International conference on teaching statistics (ICOTS9, July 2014) Falgstaff, Arizona, USA
8. Trafimow D, Marks M (2015) Editorial. Basic and Applied Social Psychology 37: 1-2.
9. Wassenstein R (2015) ASA comment on a journal's ban on null hypothesis statistical testing. http://community.amstat.org/blogs/ronald-wasserstein/2015/02/26/asa-comment-on-a-journals-ban-on-null-hypothesis-statistical-testing.
10. Diggle P, Senn S, Gelman A, Cumming G, Grant R (2014) Journal's ban on null hypothesis significance testing: reactions from the statistical arena.
11. Gaus W, Muche R (2014) Medizinische Statistik. Angewandte Biometrie für Ärzte und Gesundheitsberufe (Medical statistics. Applied biometry for physicians and health care professionals). Schattauer Verlag, Stuttgart.
12. Miller RG (1981) Simultaneous Statistical Inference (2nd Edn) Springer, New-York, USA.
13. Bretz F, Schmidli H, König F, Racine A, Maurer W (2006) Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biometrical Journal 48: 623-634.
14. Stallard N (2010) A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. Stat Med 29: 959-971.
15. Dmitrienko A, Tamhane AC, Bretz F (2010) Multiple testing problems in pharmaceutical statistics. Chapman&Hall/CRC Biostatistics Series.
16. Hochberg Y, Tamhane AC (1987) Multiple comparison procedures: John Wiley & Sons, New-York, USA.
17. European Medical Agency (EMA) (2002) Points to consider on multiplicity issues in clinical trials. London, 19 September.
18. Bretz F, Westfall PH (2014) Multiplicity and replicability: two sides of the same coin. Pharm Stat 13: 343-344.
19. Ring M, Eskofier BM (2015) Data mining in the U.S. National Toxicology Program (NTP) database reveals a potential bias regarding liver tumors in rodents irrespective of the test agent. PLoS One 10: e0116488.
20. National Toxicology Program (NTP) Toxicology and carcinogenesis studies of Ginkgo biloba extract (CAS No. 90045-36-6) in F344/N rats and B6C3F1/N mice (gavage studies) (2013). Natl Toxicol Program Tech Rep Ser 578: 1-183.
21. Gaus W, Mayer B, Muche R (2014) Sind Meta-Analysen orientierend oder konfirmatorisch? Einige Überlegungen. (Are meta-analyses exploratory or confirmative? – Some thoughts). Poster at the annual meeting of the German Region of the International Biometric Society in Bremen.
22. Heinonen T, Gaus W (2015) Cross matching observations on toxicological and clinical data for the assessment of tolerability and safety of Ginkgo biloba leaf extract. Toxicology 327: 95-115.
23. Schulz KF, Altman DG, Moher D (2010) for the CONSORT Group CONSORT-Statement: updated guidelines for reporting parallel group randomised trials. BMJ 340.