REVIEW ARTICLE

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

Anja Victor, Amelie Elsäßer, Gerhard Hommel, Maria Blettner

SUMMARY

Background: When reading reports of medical research findings, one is usually confronted with p-values. Publications typically contain not just one p-value, but an abundance of them, mostly accompanied by the word "significant." This article is intended to help readers understand the problem of multiple p-values and how to deal with it.

Methods: When multiple p-values appear in a single study, this is usually a problem of multiple testing. A number of valid approaches are presented for dealing with the problem. This article is based on classical statistical methods as presented in many textbooks and on selected specialized literature.

Results: Conclusions from publications with many "significant" results should be judged with caution if the authors have not taken adequate steps to correct for multiple testing. Researchers should define the goal of their study clearly at the outset and, if possible, define a single primary endpoint a priori. If the study is of an exploratory or hypothesis-generating nature, it should be clearly stated that any positive results might be due to chance and will need to be confirmed in further targeted studies.

Conclusions: It is recommended that the word "significant" be used and interpreted with care. Readers should assess articles critically with regard to the problem of multiple testing. Authors should state the number of tests that were performed. Scientific articles should be judged on their scientific merit rather than by the number of times they contain the word "significant."

Cite this as: Dtsch Arztebl Int 2009; 107(4): 50–6
DOI: 10.3238/arztebl.2009.0050

Institut für medizinische Biometrie, Epidemiologie und Informatik Mainz: Dr. rer. physiol. Victor, Dipl.-Stat. Elsäßer, Prof. Dr. rer. nat. Hommel, Prof. Dr. rer. nat. Blettner

Authors of medical publications like to support their conclusions with p-values and with the word "significant." How should we evaluate these p-values and the frequent use of "significant"? We will start by explaining what a p-value is and what the word "significant" actually signifies.

In general, every study should be based on a hypothesis. It is not possible in practice to test this hypothesis on every relevant subject. What happens is that a medical hypothesis—for example, that a new drug is more effective than the standard treatment in reducing systolic blood pressure after 16 weeks of treatment—is only tested on a group of patients who are selected to be as typical as possible. The results from this sample are used to decide about the validity of the hypothesis. Even if it is decided that the hypothesis is correct, there is still the possibility of a mistake (probability of error), as only a sample was studied. This decision could by chance be exactly opposite to the facts. The maximal tolerable probability of this error, the so-called level of significance (α), is normally specified as 5%. The procedure which leads either to the confirmation of the hypothesis (with maximal probability of error α) or to its lack of confirmation, on the basis of the results for the sample, is the statistical test. This provides the so-called p-value as result; the decision is then made by comparing the p-value with the level of significance. If the p-value is under or equal to the level of significance, the hypothesis is regarded as established, with the maximal probability of error α. If the p-value is above the level of significance, the hypothesis is regarded as not having been confirmed *(Box 1)*.

Most publications do not restrict themselves to a single hypothesis. Instead, several hypotheses are tested on the same sample. In the above example, it might be that not only the reduction in blood pressure after 16 weeks is compared for the standard and the new preparation, but that the following comparisons are also made:

- Reduction in blood pressure after 4 and 8 weeks
- Changes in diastolic blood pressure and in blood lipids
- The study includes another two preparations, as well as a placebo. This increases the number of possible comparisons between the drugs to 10

- So-called subgroup analyses are performed subsequently. What are the results when male and female patients or older and younger patients are considered separately?

This example shows that many different tests may be performed in a study.

What happens when several different hypotheses are tested on the same group at the same time? The probability of reaching a false conclusion increases with the number of tests performed, as an error can occur in each test. If the p-value of each test is still compared with α, there is a dramatic increase in the probability that at least one of the tests contains a false conclusion. For independent tests, it is easy to calculate this overall probability of error *(Box 2)*. If there are only 20 tests of which the p-values are compared with α = 5%, it can be expected that one p-value will be under α just by chance.

The problem of multiple testing is particularly frequent and important in genetic and prognostic studies.

Genetic association analyses are studies to establish whether a disease is linked to genetic markers, such as, for example, single nucleotide polymorphisms [SNPs]. These studies usually investigate not only a single genetic marker, but a whole series at the same time. In association studies for the whole genome (1–3), markers representing the whole genome are investigated and the number of investigated markers may lie in the thousands. Similar considerations apply to gene expression analyses, in which several thousand genes are tested on a microarray. If 1000 tests are performed, each with α = 0.05, it can be expected that 50 p-values will be less than 0.05 purely by chance, in expectation leading to 50 false positive conclusions. It has in fact been

BOX 1

The statistical test and the p-value *(see also Table 1)*

1. Statement of hypotheses

Hypothesis: The drug is superior to the previous standard therapy with respect to the reduction in systolic blood pressure after 16 weeks of treatment.

Corresponding null hypothesis: The drug is not superior to the previous standard therapy with respect to the reduction in systolic blood pressure after 16 weeks of treatment.

The null hypothesis is the opposite of the hypothesis. The hypothesis is sometimes referred to as the alternative hypothesis or the counter hypothesis in statistical nomenclature.

2. Collection of data for the patient sample

3. Statistical test to evaluate the sample data

A statistical test assumes that the null hypothesis is true and tests whether, under this assumption, the values measured for the sample are plausible (In this example: The values with the new drug are not better than those with the standard preparation) or rather implausible (In this example: The values with the new drug are clearly better than those with the standard preparation). A statistical test is used to calculate a plausibility parameter from the sample data (How probable is the result if the null hypothesis is correct?). This is presented as a probability between 0 and 1 and is known as the p-value. The more improbable the observed data are if the null hypothesis is correct, the more convincing is the evidence against the null hypothesis and for the hypothesis, and the lower is the p-value.

4. Test decision: Can the hypothesis be accepted?

If the p-value is small, the observed data are improbable if the null hypothesis is valid. This is evidence against the validity of the null hypothesis. Thus, a small p-value speaks for the opposite, the hypothesis. If the p-value is less than a prescribed limit α (the level of significance), the hypothesis is accepted. The problem is that the sample may contain random unusual values, even if the null hypothesis is really correct. If we accept the hypothesis, although it is false, we make an error. This is known as a type I error. The probability of a type I error is limited with the level of significance of the test, α. This is referred to as a statistical test at level α. Generally, α is set at 5%. This means that only in 5% of cases will the hypothesis be wrongly accepted. The test decision is made by comparing the calculated p-value with the prescribed α. The result of a statistical test is described as significant when the calculated p-value is smaller than or equal to the prescribed α. In this case, the hypothesis is accepted with the maximal probability of a type I error of α.

If the p-value is larger than the level α, the hypothesis cannot be accepted. It is, however, incorrect in this case to assume that the null hypothesis is valid, as type II errors cannot be controlled for—in contrast to errors of the type I errors. In this example, the type II error describes the error that the new drug is in fact better, but the null hypothesis is nevertheless retained. The so-called power is one minus the type II error, i.e. the probability that the hypothesis is rightly accepted. For the study to be successful, the power must be as large as possible. However, the power cannot be controlled once the data have been collected.

established that many conclusions from genetic association studies are not reproducible and are therefore very probably false positive results (4, 5).

In prognostic studies, many potential factors are often investigated. For example, a prognostic study on breast cancer not only included classical factors, but also numerous histological tumor properties. In studies on the prognosis of coronary heart disease, an immense number of laboratory markers are often included in addition to classical markers.

However, multiple testing also arises in many other areas, as a result of multiple endpoints, subgroup analyses, the comparison of several groups, or from interim analyses in sequential study designs.

This article is based on classical statistical methods, as described in many textbooks, as well as on selected technical publications.

Methods of multiple testing

To stem the flood of false positive results in medical research, measures are needed to control the probability of error in relation to all tested hypotheses.

Instead of only considering the level of each individual test, the familywise error rate (FWER) has been defined. This describes the probability that at least one of the tested null hypotheses is wrongly rejected. If this overall probability is controlled with a low value (for example, $\alpha = 5\%$), one can be fairly certain (95% certain with overall $\alpha = 5\%$) that no false positive conclusion is reached. Control of the FWER is described as "multiple level α", to make it clear that the probability of error applies to all tests simultaneously. In contrast, the "local level" means that the overall error is not being considered.

How do we control the FWER? Instead of comparing each p-value with the multiple level α, one has to set a lower level for each individual p-value. There are numerous procedures for selecting this lower limit. The converse procedure may also be employed. The p-value in these procedures is increased (adjusted) and then compared with the multiple level α. The advantage of adjusted p-values is that they are easier to understand for the reader, as he or she can compare the adjusted p-values as usual with α (for example, = 5%). This prevents the reader from wondering why "such a small" p-value is still not significant.

The best known method to control the FWER is the Bonferroni test. To achieve the aim that the overall error (the probability of making at least one false positive conclusion; the FWER) does not exceed α (for example, 5%), one divides the multiple level by the number of tests performed and compares each p-value with this lower limit. For example, if the number of investigated hypotheses is 100 and the selected multiple level is 5%, the p-value of each test (each hypothesis) is to be compared with 5%/100 = 0.0005. If this procedure is used with a FWER of 5%, only those hypotheses can be accepted and said to be significant if their p-values are equal to or less than 0.0005. *Box 3b* shows a calculated example. *Box 3a* contains more information on this procedure. This procedure maintains the selected level of FWER for all forms of dependency between the hypotheses. On the other hand, it is very strict, meaning that results may be overlooked.

TABLE 1

Decision based on the sample

	Retain null hypothesis	Accept hypothesis
Null hypothesis is correct	Correct decision	Type I error
Hypothesis is correct	Type II error	Correct decision

The Bonferroni-Holm procedure is a modification of the Bonferroni procedure, but with increased power. This involves sorting all p-values according to size and comparing them with increasing limits *(Box 3a, Box 3b)*.

Another possibility to control the FWER is to use the principle of hierarchical ranking. This means that the hypotheses are ranked according to their importance before the start of the study ("a priori"). The corresponding p-values are then compared with the selected multiple FWER level, following this sequence and starting with the most important hypothesis. Hypotheses can then be rejected in decreasing order of importance, until for the first time the p-value is not smaller than the selected multiple level (FWER). This procedure offers the advantage that all p-values can be compared to the full level (for example, 5%). On the other hand, once the level has been exceeded, no

BOX 2

Probability of falsely rejecting at least one null hypothesis (= falsely accepting a hypothesis = postulating a false positive result), if 10 independent tests are performed at the local level of 5%

= 1 – probability that no null hypothesis in all ten tests is falsely rejected

= 1 – (probability of no false rejection for each test)^{10}

= 1 – (1 – probability of a false rejection for each test)^{10}

$= 1 - (1 - \alpha)^{10}$

$= 1 - (0.95)^{10}$

= 1 – 0.60

= 0.4 = 40%

further hypothesis can be accepted, whatever the size of all subsequent p-values, even if many of them are much smaller than the level (*Boxes 3a and 3b*). This procedure is particularly suitable for clinical studies with clearly ranked main endpoints—for example, efficacy as the most important hypothesis, followed by a lower rate of adverse effects as the second hypothesis. This procedure is unsuitable for exploratory studies (such as genetic studies), for which a priori ranking of the hypotheses is impossible.

We will only point out frequent errors in the application of two other procedures often used to control the FWER. Fisher's LSD test only controls the FWER when a maximum of three groups are compared in pairs. If more than three groups are compared, this test is not an adequate method to control the FWER. The Dunnett procedure is often used to compare different dosages against a control. This procedure is only valid for the comparison with the control. It is not valid for comparisons between the different dosages.

For more information on these and other procedures, we refer you to the book of Horn and Vollandt (6) (in German language).

The probability of wrongly rejecting at least one hypothesis (the FWER) rapidly increases with the number of tests. As can be seen in the above examples, strict rejection criteria must be fulfilled to control the FWER. Thus if the procedure for multiple testing is rigorously applied to a study with many tests, this may lead to lower statistical power. In other words, valid conclusions are overlooked. This is often wrongly interpreted as a negative proof. In studies with additional follow-ups, it may be important to miss as few as possible potential leads, even if this implies accepting some erroneously significant hypotheses. For such situations, it is possible to use the false discovery rate (FDR) as a less strict possibility of controlling errors. This definition controls the expected proportion of wrongly rejected hypotheses relative to all rejected hypotheses *(Table 2)*.

BOX 3a

Bonferroni-Holm and explorative Simes (Benjamini-Hochberg) procedures and adjustment of p-values in the Bonferroni procedure

Bonferroni procedure with adjustment of the p-values

Analogously to the Bonferroni procedure as described in the text, it is also possible to adjust the p-values. This is achieved by multiplying the p-values by the number of hypotheses. If this adjustment gives a value above 1 (for example with 30 tests and a p-value of 0.04 : 30 × 0.04 = 1.2), the adjusted p-value is set at 1, as p-values (as probabilities) may not exceed 1. The adjusted p-values are then compared with the overall level α.

Bonferroni-Holm procedure

First all p-values are sorted by size and then compared with increasing limits. As in the Bonferroni procedure, the lowest limit is the overall limit divided by the number of hypotheses. The level for the next p-value is then the overall limit divided by the number of hypotheses minus 1. The limit for the third p-value is then the overall level divided by the number of hypotheses minus 2, etc.. In an example with the level 5% and 100 hypotheses:

- The smallest p-value is to be compared to 5%/100 = 0.0005.
- The second smallest p-value is to be compared to 5%/99 or ca. 0.000505.
- The third smallest p-value is to be compared to 5%/98 or ca. 0.00051 etc..

Null hypotheses can be rejected if the corresponding p-values are less than the corresponding limit. However, this only applies till the limit is exceeded for the first time. This procedure also controls the familywise error rate (FWER) for all forms of dependency between the hypotheses.

Explorative Simes procedure / Benjamini-Hochberg procedure

For this procedure too, the p-values must be sorted by size. The smallest p-value must then be compared with the Bonferroni limit – the selected false discovery rate (FDR) level, divided by the number of hypotheses. The second smallest p-value must be compared with the level multiplied by 2, divided by the number of hypotheses. The third smallest p-value must be compared with the level multiplied by 3, divided by the number of hypotheses, etc.. For example, with the selected FDR level of 5% and 100 hypotheses, the levels increase as follows:

- The smallest p-value is to be compared with 5%/100 = 0.0005
- The second smallest p-value is to be compared with 5% × 2/100 = 0.001
- The third smallest p-value is to be compared with 5% × 3/100 = 0.0015, etc..

In contrast to the Bonferroni-Holm procedure, this procedure is not restricted to the rejection of null hypotheses up to the first time that the limit is exceeded. This procedure allows for the rejection of all null hypotheses with a p-value smaller than the largest p-value which lies under the corresponding limit. In the case of independence and of so-called positive regression dependency (PRDS, a special form of positive dependency) of the hypotheses, this procedure controls the FDR at the selected level.

BOX 3b

Example of the use of the presented multiple test procedures

Four hypotheses are tested with a familywise error rate (FWER) of 5%, or with a false discovery rate (FDR) of 5%.
The resulting p-values are $p_1 = 0.03$, $p_2 = 0.01$, $p_3 = 0.035$, and $p_4 = 0.30$.

Procedure for the Bonferroni correction (control of the FWER)

Comparison of the p-values with 5%/number of tests = 5%/4 = 1.25% = 0.0125. Only p_2 is less than this limit and only the corresponding hypothesis can be designated as significant. Conversely, the p-values can also be adjusted by multiplying by 4, i.e. the number of tests. This gives the adjusted values of adj. $p_1 = 0.03 \times 4 = 0.12$, adj. $p_2 = 0.01 \times 4 = 0.04$, adj. $p_3 = 0.035 \times 4 = 0.14$, and adj. $p_4 = 0.30 \times 4 = 1.2$. The latter value is greater than 1 and is therefore set at adj. $p_4 = 1$. The adjusted p values can then be compared with the overall limit of 5%, which gives the same result as with the adjusted limits.

Procedure for the Bonferroni-Holm correction (control of the FWER)

Firstly, the p-values must be sorted according to size:
$p_2 = 0.01$; $p_1 = 0.03$; $p_3 = 0.035$; $p_4 = 0.30$
The limits in increasing order are: 5%/4 = 0.0125; 5%/3 = 0.0167; 5%/2 = 0.025; 5%/1 = 5%.
Comparison of the smallest p-value with the lowest limit: $p_2 = 0.01<0.0125$: The corresponding null hypothesis can be rejected.
Comparison of the second smallest p-value with the second limit: $p_1 = 0.03>0.0167$: End of the procedure. No further null hypothesis can be rejected and no additional hypothesis is acceptable.

Procedure with hierarchical ranking (control of the FWER)

First case: The hypotheses were ranked as follows: The most important hypothesis was H_1 (corresponding to p_1), then H_2, then H_3 and lastly H_4.
As $p_1 \leq 0.05$, the result for H_1 can be described as significant. The same applies to H_2 and H_3, but not to H_4, as $p_4>0.05$.
Second case: The hypotheses were ranked as follows: The most important hypothesis was H_4 (corresponding to p_4), then H_3, then H_2 and lastly H_1.
As $p_4>0.05$, H_4 cannot be described as significant. The same applies to all other hypotheses, as the p-value of the highest ranked hypothesis in the hierarchy was too large.

Procedure with the exploratory Simes procedure (also known as the Benjamini-Hochberg procedure; control of the FDR)

Firstly, the p-values must be sorted according to size:
$p_2 = 0.01$; $p_1 = 0.03$; $p_3 = 0.035$; $p_4 = 0.30$
The limits in increasing order are: 5%/4 = 0.0125; 5%/4 x 2 = 0.025; 5%/4 x 3 = 0.0375; 5%/4 x 4 = 5%.
Comparison of the smallest p-value with the lowest limit $p_2 = 0.01<0.0125$: corresponding hypothesis acceptable.
Comparison of the second smallest p-value with the second limit $p_1 = 0.03>0.025$: procedure nevertheless not yet ended.
Comparison of the third smallest p-value with the third limit $p_3 = 0.035<0.0375$: thus the null hypothesis corresponding to p_3 can be rejected. In addition, all hypotheses with smaller p-values are rejected. This includes not only the hypothesis for p_2, but also the hypothesis for p_1, even though this failed to reach its own limit.
Comparison of the largest p-value with the fourth limit $p_4 = 0.30>0.05$: no further rejection possible.

Summary of the results in this example: rejected null hypotheses = significant hypotheses

- With the Bonferroni procedure: H_2
- With the Bonferroni-Holm procedure: H_2
- With hierarchical ranking as in the first case: H_1, H_2, H_3
- With hierarchical ranking as in the second case: none
- With the explorative Simes procedure (Benjamini-Hochberg): H_1, H_2, H_3

In spite of the wider limits, there are no more rejections in this example with the Bonferroni-Holm procedure than with the Bonferroni procedure. However, there are more rejections with control of the FDR. Both the benefits and risks of hierarchical ranking are clear (in comparison with a priori ranking in case 1 or case 2). The ranking must therefore always be for pertinent reasons.

The most widely used procedure to control the FDR is the so-called explorative Simes procedure, referred to by most authors as the Benjamini-Hochberg procedure. This procedure was mentioned by Simes (7), although Benjamini and Hochberg (8) were the first to show that this procedure controls FDR. The exact procedure is described in *Box 3a* and *Box 3b* shows a calculated example. This procedure leads to more rejections than the Bonferroni-Holm procedure. The FDR control employs a less strict error criterion. This increases the power, although more false positive conclusions are accepted. The FDR should therefore not be used as error definition in clinical studies, but in more exploratory investigations.

In general, to circumvent or minimize the problem of multiple testing, particularly in clinical studies, one should select one or very few principle hypotheses, which are then tested for confirmation using a procedure to control the FWER. All other tests performed may not be designated with the word "significant" and must be interpreted with care. This procedure has also been suggested by the European Medicines Agency (EMEA) (9). For purely exploratory studies, mainly intended to generate hypotheses, either the FDR can be used as the definition of error, or the correction for multiple testing can be dispensed with. In the latter case, no reference may be made to significant results, but only to strikingly small p-values, which might serve to encourage additional and perhaps confirmatory studies. It must be made clear that these may be chance results, as there had been no control of any sort for the probability of type I error.

Results

As test methods are now highly automated and it is possible to record countless data for each patient (laboratory values and genetic data, etc.), very many tests are performed in each study. If the problem of multiple testing is ignored, this leads to numerous false positive findings, which are then published. Once false positive findings have been published, it lasts a long time before these are disproved and even longer before this is generally known. One must be clear that the word "significant" is often wrongly used and is in no way a criterion for quality. If the problem of multiple testing is ignored, the word "significant" has forfeited its meaning of a limited probability of error; a result wrongly described as "significant" can be totally worthless for the interpretation.

It is therefore absolutely essential that research scientists plan their studies well and, if possible, only select one (or a few) main endpoints. Articles must contain an honest statement of the number of tests performed and use appropriate procedures to decide on "significance." It must be regarded as data manipulation if many tests are performed, then all p-values are compared with α, but only those results are mentioned for which p≤α and these are then described as significant.

TABLE 2

Explanation of the error rates: The FWER is the probability that V > 0; the FDR is the expected value of (V/R).

	Null hypotheses retained	Null hypotheses rejected	
True null hypotheses	U	V	m_0
False null hypotheses	S	T	m_1
		R = V + T	$M = m_0 + m_1$

The licensing authorities also point out this problem and emphasize that it must be considered in clinical studies (9).

Discussion

In general, the reader must critically evaluate the conclusions drawn in an article (10). In particular, the problem of the lack of consideration of the issue of multiple testing is very widespread and is mostly underestimated. This opinion is based on the authors' personal experience in the statistical supervision of many medical research projects, as well as on their work as peer reviewers for medical journals. It is also supported in medical publications (11, 12). Two literature reviews have been prepared at the Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) at the University of Mainz on the association between breast cancer and polymorphisms in the COMT or SULT1A1 genes; these established that multiple tests were performed in most of the original articles. This was the case in 28 of the 34 articles on COMT and in 10 of the 14 articles on SULT1A1. However, correction for multiple testing was almost never performed. This problem was only considered in 4 of the 28 studies with several tests for COMT and in one of the 10 studies with multiple tests for SULT1A1. Thus, the problem of multiple testing was ignored in about 9 out of 10 original publications, in spite of the fact that it occurred.

The reader, the editor, and the reviewer must all take care that the term "significant" is not used inappropriately, but that the problem of multiple testing is properly considered. If the term "significant" is used copiously, suspicion is called for. Results which are presented as being "only" explorative should not be regarded as being inferior to results claimed to be significant without adequate measures to handle multiple testing. A result which is wrongly stated to be "significant" is inferior to a result which is rightly interpreted with care. It is then a mistake if the reader assumes that the word "significant" implies that the probability of error is controlled. It is unfortunately not always possible to recognize multiple testing. If an author conceals the fact that he has actually performed very many tests and has published only his most striking result, the reader is not in the position of being able to evaluate the result in the context of the number of tests performed. The reader has to see if there are signs that more tests were performed than those listed. For example, the authors may refer to other publications (including their own), in which the group of patients or the study has already been described.

Even when the authors mention that they have used methods to consider multiple testing, it is difficult for a reader without statistical training to decide whether the method used has solved the problem correctly, as there are many possible different methods. For this reason, we mentioned conventional simple methods in the methods section, together with frequent errors. In general, results from a study with many tests—such as genetic association studies or prognosis studies—should only be regarded as probably correct once they have been independently reproduced.

Conflict of interest statement
The authors declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 9 February 2009, revised version accepted on 21 July 2009.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Sladek R, Rocheleau G, Rung J, et al.: A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007; 454: 881–5.
2. Samani NJ, Erdmann J, Hall AS, et al.: Genome-wide association analysis of coronary artery disease. NEJM 2007; 357: 443–53.
3. The Wellcome Trust Case Control Consortium: A genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. Nature 2007; 447: 661–78.
4. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG: Replication validity of genetic association studies. Nature Genetics 2001; 29: 306–9.
5. Lohmüller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics 2003; 33: 177–82.
6. Horn M, Vollandt R: Multiple Tests und Auswahlverfahren. Stuttgart 1995: Gustav Fischer Verlag.
7. Simes RJ: An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73: 751–4.
8. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistic Society 1995; 57: 298–300.
9. EMEA: Points to consider on multiplicity issues in clinical trials, www.emea.europa.eu/pdfs/human/ewp/090899en.pdf
10. Prel JB du, Röhrig B, Blettner M: Critical Appraisal of Scientific Articles—Part 1 of a Series on Evaluation of Scientific Publications. Dtsch Arztebl Int 2009; 106(7): 100–5.
11. Cardon LR, Bell JI: Association study designs for complex diseases. Nature Reviews Genetics 2001; 2: 91–9.
12. Risch N: Searching for genetic determinants in the new millenium. Nature 2000; 405: 847–56.
13. Traunecker KB: Assoziation der genetischen Polymorphismen am Beispiel von SULT1A1 und COMT mit Brustkrebs. Promotionsschrift am Fachbereich Medizin der Universitätsmedizin der Johannes-Gutenberg Universität Mainz.

Corresponding author
Dr. rer. physiol. Anja Victor
Institut für medizinische Biometrie, Epidemiologie und Informatik
Universitätsmedizin der Johannes-Gutenberg-Universität Mainz
Obere Zahlbacher Str. 69
55101 Mainz, Germany
victor@imbei.uni-mainz.de