Multiple testing: when is many too much?

Rolf H H Groenwold[1,2], **Jelle J Goeman**[2], **Saskia Le Cessie**[1,2] and **Olaf M Dekkers**[1,3]

[1]Department of Clinical Epidemiology, [2]Department of Biomedical Data Sciences, and [3]Department of Endocrinology, Leiden University Medical Center, Leiden, the Netherlands

Correspondence should be addressed to R H H Groenwold
Email
R.H.H.Groenwold@lumc.nl

Abstract

In almost all medical research, more than a single hypothesis is being tested or more than a single relation is being estimated. Testing multiple hypotheses increases the risk of drawing a false-positive conclusion. We briefly discuss this phenomenon, which is often called multiple testing. Also, methods to mitigate the risk of false-positive conclusions are discussed.

European Journal of Endocrinology
(2021) **184**, E11–E14

Introduction

Is having a high level of growth hormone a risk factor for developing breast cancer? Or a high level of testosterone? Cortisol perhaps, or thyroid hormone? And if not breast cancer, pancreatic cancer perhaps? The more hormones you investigate the more likely it becomes that – just by chance – at least one of them will appear to be a risk factor, even if it is not. It is well-known that many relations found in medical research are false-positive signals (1): chance findings that do not stand up to replication. Here, we discuss the probability of incurring a false-positive finding, how this is related to the number of hypotheses that are being tested, and how to control the chance of a false-positive finding.

Multiple testing

'The more you look, the more you see'. This also applies to scientific research, particularly to hypothesis testing. In medical research, it is common to test a so-called null-hypothesis and rejecting it when the *P* value of the test is smaller than a predefined threshold. In clinical research, this threshold is mostly set at a statistical significance level of 0.05. Suppose researchers are interested in the question whether high growth hormone (GH) levels cause breast cancer. Their investigation results in a *P* value <0.05 and the researchers claim that a significant effect exists (in this case, they claim that high GH levels cause breast cancer). However, a *P* value smaller than 0.05 does not imply that we have proven with certainty that an effect exists (by the way, this also applies to a 95% CI that does not contain the null-value) (2). If the null-hypothesis is true (e.g. no effect of GH on breast cancer), there still is a probability of 5% that the null-hypothesis is falsely rejected, a so-called *type I error*. This means that, even when GH does not increase breast cancer risk, there is a 5% chance that a study using null-hypothesis testing will conclude that it does (here we assume no bias).

Suppose researchers not only investigate GH, but also testosterone, and that both of these hormones have no true relation with breast cancer. For each hormone, the researchers would again incur a 5% chance of incorrectly rejecting the null-hypothesis, and thus potentially claiming an effect that is not actually there. The probability that *at least one* of the two relations mistakenly shows 'significance' in the study is between 5 and 10%, depending on the correlation between the two hormone measurements. In most cases, the probability of at least one false-positive result is almost doubled to 10%. With more than two null-hypotheses, the probability of a false-positive result grows rapidly (Fig. 1).

One solution to reduce the probability of false-positive conclusions is to simply reduce the number of hypotheses being tested, by specifying before analyzing

https://eje.bioscientifica.com
https://doi.org/10.1530/EJE-20-1375

Published by Bioscientifica Ltd.

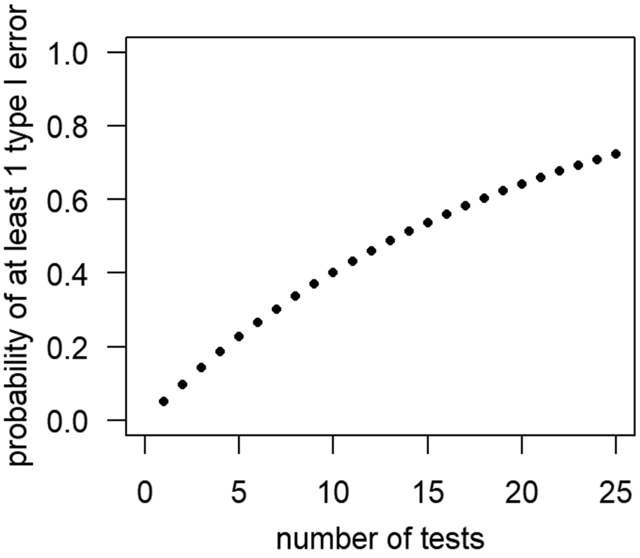

Figure 1
Relation between the number of independent hypothesis tests performed at a significance level of 0.05 and the probability of at least one true hypothesis being rejected.

the data which hypothesis is most important and refraining from testing less important hypotheses. When the number of hypotheses that require testing is still large, the probability of false-positive findings can be reduced by statistical corrections. An easy way to preserve the overall type I error (aiming for an overall 5% probability of a false-positive finding within a study), is to divide the significance level by the number of tests that are performed. This is commonly known as a *Bonferroni correction*. For two tests this means that the significance level would be 0.05/2=0.025, instead of 0.05, while for 25 tests it would be 0.05/25=0.002. Note that even if researchers did not apply a correction for multiple testing, a reader could easily apply the Bonferroni correction, provided it is reported how many hypotheses were tested.

A trade-off between false-positives and false-negatives

Obviously, corrections for multiple testing reduce the number of false-positive conclusions. There is, however, an immediate consequence of applying a Bonferroni correction: the probability that a true relation may go unnoticed will increase (we call this a *type 2 error*). The probability to detect an existing true effect is commonly known as the *power of the study*. In a study in which a single test has a power of 80%, the power of each of 25 Bonferroni corrected tests would be less than 39%. This is intuitively clear, as the significance boundary is now much lower. Figure 2 shows how the power of each test goes down as the number of Bonferroni corrected tests increases.

For this reason, many researchers may be hesitant to perform a Bonferroni correction, fearful that interesting findings may be disregarded. Indeed, the Bonferroni correction can be conservative (too strict), when the data about the different hypotheses are strongly positively correlated. For example, if high glucose levels are mistakenly shown to be related to hip fractures, this will likely also be the case for HbA1c; effectively less than two tests are performed, so the probability of making at least one mistake is much less than two times 5%.

Let us consider a study performing 100 statistical tests (a number not unusual for clinical studies). Figure 3 shows the 15 smallest *P* values of this hypothetical study in which 100 different null-hypotheses are tested, of which 90 null-hypotheses are true (i.e. there truly is no relation), and 10 times it is not (i.e. there truly is a relation). If no correction for multiple testing is made, 13 out of 100 null-hypotheses would be rejected, of which four rejections are wrong, because the null-hypotheses are actually true (hence, false-positive conclusions are made). Also, one false null-hypothesis would not be rejected (false-negative). Applying a Bonferroni correction reduces false-positive conclusions but reduces the power as well.

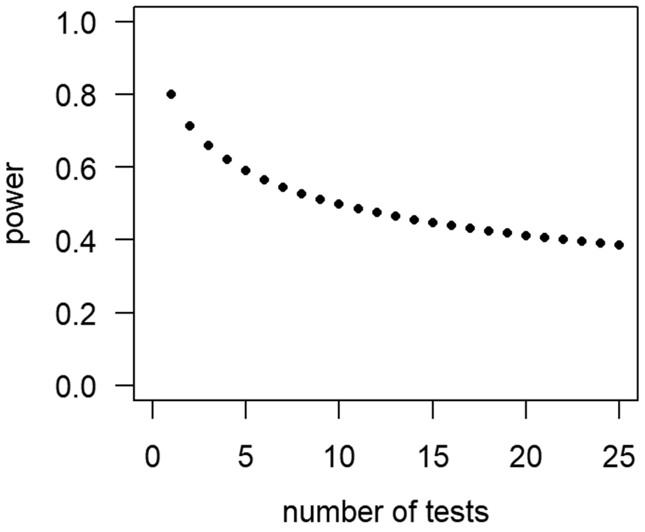

Figure 2
Relation between the number of independent hypothesis tests performed and the power of each test after Bonferroni correction. *It is assumed that, without Bonferroni correction, each test has a power of 0.8.*

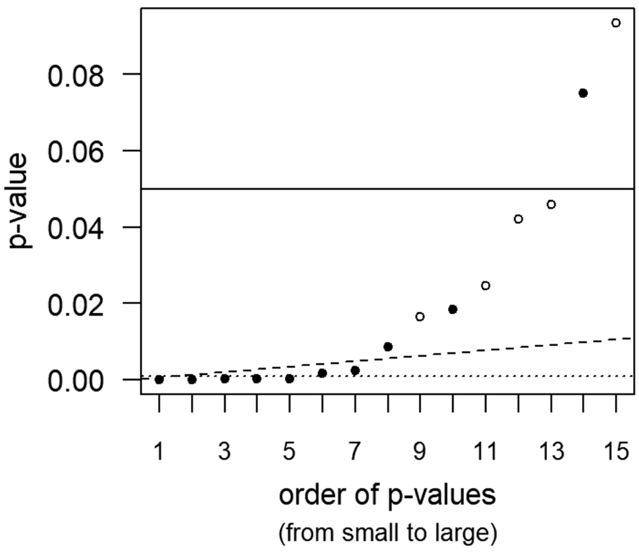

Figure 3
Illustration of methods to correct for multiple testing. The smallest 15 out of 100 *P* values are presented. Open circles represent null-hypotheses that are true (no effect exists), solid dots null-hypotheses that are false (there exists an effect). The solid, dotted, and dashed lines represent the thresholds used for conventional statistical significance testing (no correction for multiple testing), used after Bonferroni correction, and used in the Benjamini–Hochberg method, respectively.

Using the more stringent Bonferroni correction in the hypothetical study leads to zero false-positive conclusions, yet five false-negative conclusions.

False discovery rate

To maintain power, while still limiting the number of false-positive conclusions, the concept of the false discovery rate (FDR) was proposed by Benjamini and Hochberg (3). FDR procedures aim for controlling the proportion of false-positive conclusions, instead of controlling the probability of at least one false-positive conclusion (as, e.g. Bonferroni correction does). A researcher who performs a Bonferroni correction in each paper that she writes, may expect to have false-positive results in at most 5% of her papers. A researcher who always uses FDR, however, could have false-positive results in every paper, but has the guarantee that on average these false-positive results comprise no more than 5% of all positive results obtained. The idea behind FDR procedures is that if many hypotheses are tested, and therefore many findings are presented in one paper, it is not so important that we can guarantee that every finding is correct, as long as the majority of findings is. FDR procedures are, for example, used in gene expression studies, in which hypotheses are tested for tens of thousands of genes and Bonferroni correction would reduce the already limited power too much. Results obtained using FDR are less certain than those using Bonferroni and therefore require follow-up studies for confirmation.

The Benjamini–Hochberg method is illustrated in Fig. 3. For this method, one first orders the *P* values from small to large. The smallest *P* value is compared to the stringent Bonferroni threshold of 0.05 divided by the number of hypotheses tested, the second smallest to a much less stringent value of twice that threshold, the third smallest *P* value to three times the Bonferroni threshold, etc. The researcher now finds the largest *P* value that makes its threshold and rejects all null-hypotheses with that *P* value or smaller *P* values. In this way, the proportion of false discoveries (false-positive conclusions) can be controlled. Since the threshold becomes less and less stringent as we go from the smallest to larger *P* values, the power is larger than for Bonferroni-type methods, where all *P* values are compared against the same (stringent) threshold. Hence, when many null-hypotheses are in fact not true, FDR-type procedures will often detect more. Let us have a look at the example in Fig. 3, where Bonferroni correction resulted in zero false-positive conclusion, yet five false-negative conclusions. Instead, when the Benjamini–Hochberg method is applied, there is still no false-positive conclusion, but instead only three false-negative conclusions.

Reporting multiple testing

Let us return to the example of a study that investigates the relations between different hormones and different cancer types. It is relatively easy to recognise multiple testing, if each of the investigated relations is reported, for example, listed in a single table. Obviously, this will be much harder, if only those relations are reported that reached statistical significance. In the case of such bad practice, it will be difficult for the reader to get an idea of what the risk of false-positive results is, and whether corrections for multiple testing should have been made. Likewise, when results of different analyses of one study (e.g. different analyses based on data from one ongoing cohort study) are reported in different articles, it will likely be unclear whether results should be judged in isolation, or whether the increased risk of false-positive results due to multiple testing should be considered.

When is many too much?

Before analyzing the data, researchers can try to avoid the problem of multiple testing, for example, by making (well-considered) choices about which hypotheses to test. Alternatively, they could make a distinction between primary and secondary null-hypotheses. In that case, however, it might be unclear what to conclude when the primary null-hypothesis is not rejected, while the secondary is. The issue of multiple testing is sometimes avoided by calling analyses 'exploratory'. In that case, the results of such an analysis should indeed be clearly reported as such (and any claims of (strong) evidence for reported effects or relations avoided). Authors and readers should be aware that studies marked as exploratory have a substantial risk of containing false-positive findings.

If researchers nevertheless want to test a large number of null-hypotheses, such as in research on genes or metabolites, the risk of drawing a false-positive conclusion is inflated unless a proper correction is carried out. It is worth mentioning that also the estimates of those relations that reach statistical significance are biased, which – unfortunately – cannot easily be corrected for. We have described two fairly simple methods, Bonferroni and Benjamini–Hochberg, that can be used to correct *P* values for multiple testing. They can even be applied by the reader of an article if the authors did not perform the correction themselves, provided the number of hypotheses that were studied is transparently reported. A statistician may be consulted for more advanced and powerful methods for specific research designs, for example, when there are primary and secondary hypotheses (4), or when the data for different hypotheses are strongly correlated (5).

Declaration of interest

O M D is a Deputy Editor for the *European Journal of Endocrinology*. He was not involved in the review or editorial process for this paper, on which he is listed as an author. The other authors have nothing to disclose.

Funding

This work was supported by grants from the Netherlands Organization for Scientific Research (ZonMW-Vidi project 917.16.430) and the LUMC.

References

1 Ioannidis JP. Why most published research findings are false. *PLoS Medicine* 2005 **2** e124. (https://doi.org/10.1371/journal.pmed.0020124)

2 Dekkers OM. Why not to (over)emphasize statistical significance. *European Journal of Endocrinology* 2019 **181** E1–E2. (https://doi.org/10.1530/EJE-19-0531)

3 Benjamini Y & Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* 1995 **57** 289–300. (https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)

4 Bretz F, Maurer W, Brannath W & Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009 **28** 586–604. (https://doi.org/10.1002/sim.3495)

5 Goeman JJ & Solari A. Multiple hypothesis testing in genomics. *Statistics in Medicine* 2014 **33** 1946–1978. (https://doi.org/10.1002/sim.6082)

Received 1 December 2020
Accepted 7 December 2020