Hans-Peter Dürr and Martin Scott

AMNOG and the determination of an added benefit:

Modern visualization of results instead of mammoth dossiers!

The benefit assessment of new drugs is carried out based on a variety of statistical analyses for different endpoints, subgroups, or patient populations. This creates the problem of so-called multiple testing, which confronts both pharmaceutical companies and regulatory authorities with a statistical dilemma: How can the statistical errors (type-1 and type-2 error) be adequately controlled in multiple testing? This article describes the causes of the problem and shows that the visualization of statistical test results would help both parties involved in the process: On the one hand, the pharmaceutical entrepreneur in designing the much-vaunted 'value story' for his product, and on the other hand, the regulatory authorities in assessing it.

Problem

>>In Germany, the benefit assessment of new drugs is carried out by the Federal Joint Committee (G-BA) together with the Institute for Quality and Efficiency in Health Care (IQWiG), which defines the methodological and statistical guidelines for the assessment process. The additional benefit of a new drug is examined in clinical trials or registry studies, quantified by statistical evaluation of the study data, and finally classified into the categories of a non-quantifiable, minor, considerable, or major added benefit.

The statistical evaluations lead to a problem that has been known since the beginning of statistics: the more you search, the more likely you are to find something. This problem of multiple testing also applies to finding treatment effects regarding the efficacy of a drug. For example, excessive statistical testing can lead to the surprising result that aspirin is particularly effective for people born under the zodiac sign of Capricorn but is completely ineffective for those born under the zodiac sign of Libra or Gemini (Peto et al. 1995). Could this be a mistake based on a methodological problem? Yes - of course this is the case.

A central term in this context is the so-called type-1 error α. It indicates the probability with which an effect that is not actually present is presented as statistically significant and thus leads to a false positive conclusion. 'Statistically significant' means that it is unlikely this effect will be found by chance if it does not exist in reality. In this context, an event is generally described as 'unlikely' if it occurs - assuming randomness (the null hypothesis) - in less than 5% of cases ($\alpha<5\%$). If there is an effect in reality, and statistical analysis nevertheless postulates such effect, then this error is due to the type-1 error. The statistical approach falls victim to a systematic error in 5% of cases, according to the type-1 error.

Summary

The benefit assessment of a drug is usually based on a large number of statistical tests. However, excessive statistical testing can postulate effects, that don't exist. This not only concerns positive effects, such as the effectiveness of a drug, but also negative effects, usually called adverse effects. Methods to correct these errors exist, but - if implemented consistently - they lead to the opposite problem: the actual effect is negated, as well as the effectiveness and any adverse effects. The recommendations formulated by the G-BA and IQWiG Specifications for the number of analyses and the correction of so-called multiple testing are difficult to meet in some cases. A solution to the problem does not seem to be in sight, at the same time an available option could be better used: the use of modern methods for the visualization of results.

Keywords

AMNOG, benefit assessment, statistics, multiple testing

The consequence of multiple testing is: let us perform 20 statistical tests, and each test commits with a probability of 5% a mistake, then we will have an average of 1 test result postulate an efficacy purely by chance (with probability of 20 x 5% = 100% !). This leads to the core of the problem: For the benefit assessment of a drug, generally not only 20 statistical tests are performed, but hundreds, often thousands. This makes it likely to postulate an effect that does not exist in reality. This problem is intensified in the medical benefit assessment due to the obligatory interaction and subgroup analyses, which inflate the number of statistical analyses to be performed immensely - and thus also the type-1 error.

Subgroup characteristics that need to be examined frequently are, for example, age, sex, severity or stage of the disease, previous treatments, etc. The relevant question for patient health is whether the effectiveness of a drug depends on the subgroup characteristic. Using the subgroup characteristic 'sex' as an example, we would like to know whether the drug works better in men than in women, or vice versa. If this is the case, there is a so-called interaction. The guidelines established by the G-BA and IQWiG require subgroup characteristics to be considered in detail. Nevertheless, IQWiG states that "If several subgroups are analysed, results in a subgroup may well reach statistical significance, despite actually being random" (IQWiG 2017).

Solution approaches

There are several methods to control the problem of multiple testing (Bender et al. 2007). Common to all these methods is the lowering of the level of significance, which makes it more difficult to find desired effects. Thus, statistical tests carried out in excessive numbers ruin the probability of being able to prove the efficacy of a drug, or in other words: It is precisely the attempt to prove efficacy that worsens the chance of being able to provide this proof statistically.

In the simplest method, the so-called Bonferroni correction,

the level of significance is lowered by the fact that the type-1 error α is divided by the number of statistical tests - the more tests, the higher the requirement for the statement 'test passed'. With 100 statistical tests. the significance level should be lowered from α=5% to α=0.05%. The effectiveness of a drug would only become 'statistically significant' if the effect was very strong or if the sample size is very high.

When it comes to adverse effects, the opposite problem occurs: The lowered level of significance sets the detection threshold for adverse effects so low that in practice many adverse effects that actually occur would be described as 'statistically not significant'. The scientifically correct and consistent implementation of Bonferroni methods leads to an over-conservative rejection of effects, regarding both the effectiveness of a drug as well as the risk of adverse effects.

Already in 1998, Feinstein described the problem as a 'clinico-statistical tragedy' and called for a 'move away from multiple testing, back to clinical relevance' (Feinstein 1998). In particular, he addressed the problem that the number of statistical analyses rises sharply due to interaction and subgroup analyses and that ultimately false results are created by 'data dredging', i.e. a rather unspecific 'squeezing out' of study data. The problem he describes is more relevant than ever, even if it has become more socially acceptable after the turn of the millennium due to positive terms like 'data mining' or 'big data'.

At present, there is no generally accepted procedure that is scientifically correct and equally practicable; in the sense that, on the one hand, it is sufficiently probable that efficacy of a drug is discovered and on the other hand the risk of adverse effects is adequately excluded. However, the G-BA's document template for the preparation of dossiers published in February 2019 indicates a trend towards placing more emphasis on the visualisation of results of statistical tests (G-BA 2019), for example in the form of forest plots and survival time curves.

Outlook

The current practice of benefit assessment of drugs suffers under a methodological conflict of interest. The attempt to test and answer many detailed questions collides with the statistical problem of multiple testing. Statistical methods for correcting multiple testing (e.g. the above-mentioned Bonferroni correction) can often not be used consistently,

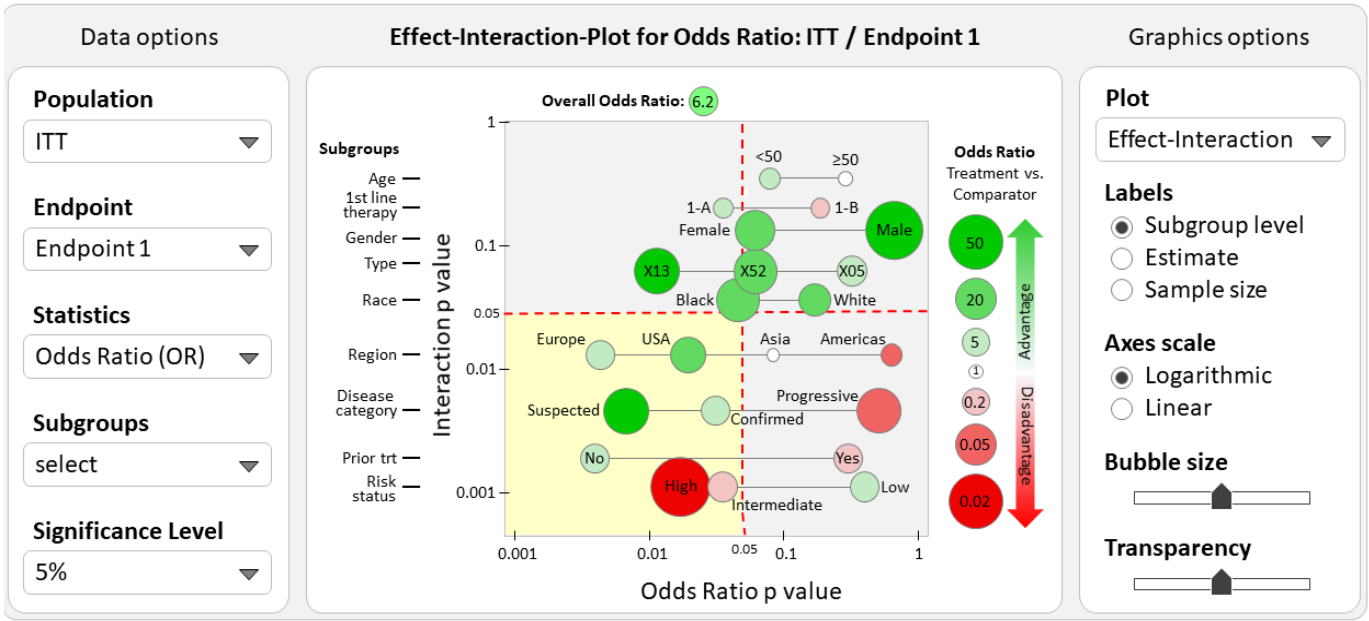

Figure 1: Effect interaction graph for the presentation of interaction and subgroup analyses in the context of benefit assessments (https://de.numerus.com/IEBfree). In this example, the treatment effect is represented by the Odds Ratio (OR) and shown for an endpoint and the ITT population. Y-axis: p-value for the interaction of the treatment effect with the subgroup characteristics (age, gender, etc.). X-axis: p-value of the treatment effect for a certain subgroup (e.g. Age: < 50 and ≥ 50). Values within a subgroup are connected by horizontal lines. Green: positive treatment effect, red: negative treatment effect. Effect sizes (here: OR) are visualized by a proportional circle diameter of the bubbles (see legend on the right). The overall effect is shown at the top of the graph (OR=6.2). The yellow coloured quadrant contains the statistically significant effects (here: α= 0.05). Data and graphic options can be easily adjusted in the left and right panel regarding the specific question. The data example shows a situation in which the main disadvantage of the drug is found in the patient subgroup with risk status 'high'. Source: Numerus, www.numerus.com.

because this would lead to both the effectiveness and harmfulness of a drug not being detected with sufficient sensitivity. Feinstein's formulation of the 'clinico-statistical tragedy' describes the extent of this dilemma in a drastic way. It seems we have entered an era in which excessive statistical testing is relegated to the background and replaced by an interactive, explorative, balanced overview of the results.

The theoretical solution to the problem presented is initially clear: structurally, a rating or scoring problem exists. The current classification of added benefits into categories minor, considerable, or major is the result of such a rating procedure. However, a unique rating algorithm cannot be used because high-dimensionality is present: A benefit assessment usually considers the dimensions population(s), endpoint(s), subgroup(s), and evaluation method(s). Each of these dimensions contains several, even dozens of characteristics or outcomes; the combinatorial matrix then quickly reaches hundreds or even thousands of individual results. Each of these individual results would have to be weighted separately because different endpoints - e.g. mortality and quality of life – usually do not have the same relevance to patients. In total, such procedures would therefore require thousands of expert opinions, and this separately for each disease. It is clear that feasibility is quickly exceeded. What can help here?

If the practical solution of the problem is so difficult, this leads back to the strategy described above of an appropriate presentation and visualization of results. A dossier of several thousand pages for the benefit assessment of a drug can claim to be 'comprehensive', but not to offer 'clarity'. It would be desirable for all parties involved to be offered statistical results for evaluation in a standardised and clear form, as shown in the figure above. This would benefit all those involved: the pharmaceutical entrepreneur could explore and present the opportunities and risks of his drug more easily and more reliably. The regulatory authorities could arrive more quickly and reliably at an overall assessment of the benefits of the treatment.

Perhaps the following formulation, which is currently widely used, summarises the problems and the potential for improvement in an appropriate way- dialogue and consensus-building at eye-level. This could mean in the case of the benefit assessment of medicinal products: Through appropriate visualization of the results of statistical analyses.<<

Literature

Bender R, Lange S and Ziegler A (2007) [Multiple testing]. Dtsch Med Wochenschr 132 Suppl 1:e26-29.

Gemeinsamer Bundesausschuss (2019) Anlage II.6: Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen. Dokumentvorlage, Version vom 21.02.2019.

Feinstein AR (1998) The problem of cogent subgroups: a clinicosta¬tistical tragedy. J Clin Epidemiol 51:297-299.

IQWiG (2017) Allgemeine Methoden: Version 5.0, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln.

Peto R, Collins R and Gray R (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 48:23-40.

PD Dr. Hans-Peter Dürr
is a health scientist and habilitated in epidemiology at the University of Tübingen. In addition to his research topics in the field of mathematical and statistical modelling in medicine, he heads the Department of Epidemiology and Scientific Methodology at Numerus and is responsible for epidemiological issues and statistical analyses in the fields of HTA/value assessment, modelling and simulation.
Contact: hans-peter.duerr@numerus.com

Martin Scott
is a statistician and managing director of Numerus GmbH Germany and works in the field of drug and medical device research in various therapeutic areas and study types, including phase IV registration studies, HEOR and HTA. He previously worked in several global pharmaceutical companies and clinical research organizations and is currently advising Market Access Teams with special regard to the requirements of AMNOG in Germany.
Contact: martin.scott@numerus.com