Multiple endpoints

asterix

A single endpoint

An endpoint (EP) is a variable that contains information on the disease-related condition of a patient. It is intended to measure the physiological functions, the well being, or the time to a disease related event. Most clinical trials use a single primary endpoint to study the efficacy of a treatment.

***Example** for a single EP*

In trial for a new medication in muscular dystrophy, the distance walked in 6 minutes may be chosen as a single primary endpoint.

Multiple endpoints

Often a single endpoint is not sufficient to cover all study goals. Some diseases are complex and a new treatment needs to have an effect on several endpoints simultaneously. In that case the endpoints are referred to as co-primary. In other diseases, a treatment may be considered beneficial even if it has an effect at least in one out of several endpoints.

***Example** for co-primary EP's*

A new radiation therapy for skin lesions may be considered superior only if it has, both, higher efficacy and a reduced pain side effect.

Aiming for efficacy in at least one EP In epileptic diseases a medication may be useful if it helps to reduce seizure frequency or seizure severity or both.

How to show efficacy

Modern medical research is subject to the principle of evidence based medicine. For a new treatment to be accepted, there must be sufficient objectively measured evidence that in the patient population the treatment will on average provide a better outcome than under placebo or under a suitable control treatment. The patients enrolled in a clinical trial represent a sample from the patient population. Properties observed in the sample provide an estimate of the corresponding properties of the population. However, the sampling of patients is subject to randomness. Hence, study results inevitably show some random variation. In a statistical hypothesis test we compare the magnitude of the observed effect to the random variation. Loosely speaking, we may conclude efficacy if the observed effect can hardly be explained by the random variation, and such a result is called statistically significant.

Example

Five patients are treated with a new drug and five patients are treated with placebo. In the active treatment group, 4 are cured, in the placebo group 3 are cured. The estimated increase in success rate is 20%. But given the small sample size it is clear that such a result may have well arisen by chance under the assumption of equal success rates.

Error rates

We want to avoid the individual burden and public health costs of patient exposure to ineffective treatments. Therefore our first concern when testing hypotheses concerning efficacy of a treatment is to avoid a false positive conclusion. We call the probability for a false positive conclusion the *type I error rate*. We construct our hypothesis tests in such a way that the type I error rate is below some limit, usually 2.5%.
Our next priority is to identify a treatment as efficacious if it really has a certain effect. We call the probability to achieve this goal the power of the hypothesis test. The power increases with increasing probability for a type I error, posing some trade-off. Also the power increases with sample size, because a larger sample contains more information. Ideally, tests are constructed such that the type I error rate matches the prespecified limit and such that the information in the data is used completely. For complex testing problems, this may not be easy to achieve, and so the development of refined testing procedures is required.

Multiple testing problem

If more than one EP is observed, the probability to observe some large effect just by randomness is increased. To prevent a high rate of false positive conclusions, our tests must be adjusted, such that overall the probability for a type I error is still less or equal the pre-specified level, e.g. 2.5%. Such an adjustment means that larger effects or larger sample sizes are needed to conclude efficacy in a particular EP. However, when judging the overall effect across all endpoints, showing just some effect (without necessarily referring to one particular EP), the power can beincreased and potentially sample size saved.

Example

Throw a dice. The probability for 6 is 16.7%. Throw two dices. The probability for at least one showing 6 is 30.6%. With three dices, the probability for at least one showing 6 is 42.1%. Similarly, in a clinical trial with many endpoints the chance to observe some extreme event is increasing with the number of endpoints.

Asterix methods

Fallback tests for co-primary endpoints In the classic co-primary EP test, the trial goal is achieved if all(unadjusted) single EP tests show a significant effect. If only some tests are significant, no conclusion on any EP can be drawn. Fallback tests for co-primary EP allow to use the full power of a co-primary EP test forthe main goal. They also allow for conclusions on efficacy in individual goals, even if only some goals meet certain significance requirements.

Optimal exact tests for multiple binary EP
In many studies binary EP are used, such as symptom relief Yes/No, or occurrence of an adverse event Yes/No. We want high power to detect an effect represented by several binary EP. Thus efficacy is concluded if the success numbers are large enough across all EP in the treatment group compared to placebo. However, different combinations may indicate overall success, e.g. high improvement in one EP or medium improvement in many EP.

The optimal exact tests provide maximal power to identify a certain set of potential outcomes, representing the assumed true effect sizes. These tests are exact, which means the type I error rate is controlled, even with small sample sizes.

Simultaneous inference for multiple EP with repeated measurements Often EP are measured more than once in each patient. Measurements are either taken at subsequent time points or under different treatment conditions. Methods for the analysis of dependent data are then required. In ASTERIX a method to provide simultaneous confidence intervals and hypothesis tests for multiple EP was extended to models allowing for repeated observations. The method gains from the additional information provided in repeated observations, at the sametime the intervals, even though approximate, are more narrow than those provided by simple multiplicity adjustments.

Possible benefits for patients

- Multiple endpoints provide more overall information than a single endpoint.
- Methods developed in ASTERIX allow us to increase the extent to which this additional information can be used.

Possible downsides

- Additional endpoints add additional noise, such that inference on a particular individual endpoint is less precise than in the single endpoint setting.
- The multiple testing strategy needs to be precisely defined in the study protocol. Post-hoc choices will not guarantee the desired properties.

Contact details

ASTERIX - Advances in Small Trials dEsign for Regulatory Innovation and eXcellence *This leaflet is developed together with the Patient Think Tank and is part of the Asterix project, funded by the EU FP-7 program. For more information, see our website: www.asterix-fp7.eu*