

Arbeiten aus dem Paul-Ehrlich-Institut, Band 94
10th International Paul-Ehrlich-Seminar, Lübeck 2002
Druck- und Verlagshaus Sperlich

Criteria for Clinical Efficacy – Readout and Monitoring of Clinical Studies

Hans-Jørgen Malling

The clinical efficacy of allergen-specific immunotherapy is often questioned in spite of solid documentation based on placebo-controlled, double-blind (PCDB) trials (1-3). The documentation, however, is rather based on quantity than quality of the relevant studies. Generally immunotherapy studies suffer from some fundamental methodological problems such as the inclusion of a small number of participants, a high frequency of withdrawals and conclusion based on unbalanced groups with respect of disease severity (mostly caused by a lack of a baseline assessment) (4). Selected important criteria in relation to the evaluation of the quality (and consequently the reliability) of published studies but even more essential in the planning of conclusive studies will be discussed in more detail.

Randomization

The process of randomization is insufficient in many clinical trials. Ideally, participants should be assigned to comparison groups on the basis of a random process characterized by unpredictability. The methods used to generate the random allocation sequence should minimize the risk of bias in group assignment. Restricted randomization is often used in trials including a smaller group of patients to achieve balance between groups in size and characteristics. Randomization could be in blocks of, e.g. 10 patients allocating five to each arm of the trial (5), or by using stratification according to predefined clinical criteria like age, gender, disease (6). Allocation to treatment groups using minimization (which by some is not considered a genuine randomization) has the advantage of making small groups closely similar in terms of several characteristics at all stages of the trial (7).

Blinding of intervention and assessment

The blinding of the trial is important to avoid bias at several stages (8). The use of placebo combined with blinding is important to evaluate the "true" treatment effect associated with the intervention. Blinding of patients and health care providers in subcutaneous immunotherapy often involve the inclusion of histamine in the placebo preparation, and although not perfect, it induced reactions imitating allergen local reactions. The blinding should also include the data analyst and the success of the blinding should be reported by asking participants to identify which treatment they were receiving.

Statistical methods

The statistical methods used for analyzing outcome data is a critical issue (9). In many clinical trials the applied statistics is not strictly appropriate for the particular situation. When planning a study the number of patients to be included in each treatment arm should be large enough to have high probability (power) to detect statistical significance. Power calculations depend on the expected magnitude of efficacy and the variability of included patients and must be applied in the planning phase of the trial. The size of the clinical effect to be evaluated is inversely related to the sample size necessary to detect it (large samples are necessary to detect small differences) (10). Inclusion of insufficient numbers of patients increases the risk of statistical type I errors (not being able to detect a factual effect) but also in the case of differences in disease severity between the evaluated groups at inclusion to result in statistical type II errors (demonstrating an effect that actually does not exist). The choice between parametric and non-parametric statistics is an almost religious discussion. In biologic science most small samples are not normally distributed. This may be circumvented partly by transformation of data, but showing statistical significant difference using non-parametric tests usually reinforces the conclusion. Most studies assess study findings in terms of their statistical significance. The P-value represents the probability that the observed results could have arisen solely by chance when the interventions did not differ. In order to improve the readout of clinical studies it is recommended to include confidence intervals giving the range of the differences or the magnitude of efficacy between the two rates, i.e., the range of uncertainty expected to include the true value. It is customary to present the 95% confidence interval.

Participants flow

The flow of participants through each stage of the study is critical for the evaluation of efficacy. Information for each group on the number of participants randomized, receiving treatment, completing the study protocol, and analyzed for outcome is important. Many studies do not give information on whether some participants did not receive treatment as allocated, were lost to follow up, or excluded from the analysis. Ideally, the primary outcome (clinical efficacy) should be based on all patients randomized (intention-to-treat analysis). In order to evaluate the "true" treatment effect, it is crucial to include participants leaving the study due to insufficient clinical efficacy

in the primary outcome. Effects should be included.

Minimal clinically important difference

In the discussion of clinical efficacy, patients suffer from symptoms and the need for treatment based on a reduction in symptom severity. A relevant reduction in the number of participants statistically significant is observed. The magnitude of symptom scores and the morbidity of the symptom/medication. The placebo effect has a statistically significant difference. In convenience, this figure

Grading system for clinical efficacy

A useful system for grading clinical efficacy is proposed. The score do not involve rate data on symptom and medication scores. A significant improvement in symptom score, versa, still represents a significant improvement. Problems arise when cases, it must be considered. It has been induced by the

Importance of pre-treatment

Most clinical studies are conducted over a period (for seasonal allergies) consuming but important for the groups and for the evaluation of efficacy by direct comparison with the treated group. In randomized controlled trials, comparable groups are allocated to the intervention and the control group. The severity of the disease in the control group is a statistical methods for

in the primary outcome analysis. Likewise, patients leaving the study due to adverse effects should be included in the safety analysis.

Minimal clinically relevant efficacy

In the discussion of the clinical efficacy it is important to remember that allergic patients suffer from the clinical manifestations of the disease that is the allergic symptoms and the need for intake of anti-allergic drugs. Consequently efficacy should be based on a reduction in these parameters. The magnitude of efficacy (minimal clinically relevant reduction in disease severity) might be debated. By including a high number of participants statistically significant, but clinically irrelevant differences might be observed. The magnitude of efficacy should be clinically relevant, i.e., the reduction in symptom scores and drug consumption should from a clinical point significantly reduce the morbidity of the disease. Based on an evaluation of 68 PCDB studies presenting symptom/medication scores (3) a reduction in disease severity of > 30% additional to the placebo effect has been considered clinically relevant (and compatible with statistical significant differences). Depending on the risks of intervention, costs, and inconvenience, this figure is subject to changes.

Grading system for clinical efficacy

A useful system for the interpretation of the value of treatment in relation to clinical efficacy is proposed. Studies providing data on a combined symptom/medication score do not involve logistic problems. The interpretation of studies presenting separate data on symptom scores and medication scores may be distorted, due to the relative weight of these parameters. Statistical significant improvement in both symptoms and medication scores > 30% indicates clinical relevant efficacy ("Efficacy"). Statistical significant improvement in symptoms > 30% and no change in drug scores, or vice versa, still represent an improvement of the total disease severity ("Possible efficacy"). Problems arise when symptoms and drug score show discordant changes. In these cases, it must be concluded that no significant and indisputable clinical efficacy has been induced by the treatment ("No efficacy").

Importance of pretreatment monitoring

Most clinical immunotherapy studies do not include a pretreatment monitoring period (for seasonal allergies a pretreatment season). The risk of excluding this time consuming but important information related both to the allocation to treatment groups and for the documentation of clinical efficacy. Many clinical studies evaluate efficacy by direct comparison of the results obtained in the placebo and the actively treated group. In relation to group assignment the conclusions are based on comparing comparable groups with respect to parameters important for the primary outcome. The allocation to the intervention groups is based on subjective data (patients recall of disease severity in the preceding season) implying a high risk of biasing the groups. The statistical methods have no change to compensate for variations in disease severity in

the different groups, and consequently there is a substantial risk for introducing both statistical type I (false rejection of an existing difference in treatment effects) and type II (false acceptance of a treatment effect caused by preexisting pretreatment difference in disease severity of the two groups). The advantage of including a pretreatment monitoring period is that the disease severity (patient scoring of clinical symptoms and use of rescue drugs) can be precisely determined based on objective parameters and the patients randomized appropriately without bias. Furthermore, looking at changes from pretreatment a further compensation for variation in disease severity scoring might be obtained. By using the patient as his own control in relation to the clinical outcome one additionally has the possibility to detect both improvements and deteriorations in disease severity, i.e., the natural outcomes of untreated (placebo treated) disease, and the magnitude of changes. This approach gives the capacity to identify specified characteristics of responders of the intervention versus non-responders, and to correlate para-clinical parameters to the magnitude of clinical improvement. A crucial factor for the approach to evaluate changes in disease severity is the constancy of allergen exposure during succeeding evaluation periods. Working with pollen allergy a substantially lower pollen count in the treatment season versus the monitoring season might make it impossible to draw conclusions.

Conclusion

When designing clinical studies evaluating the clinical efficacy of allergen-specific immunotherapy several criteria are essential:

- The study should be a placebo-controlled, double-blind, randomized trial
- The sample size should be large enough to detect statistically significant and clinical relevant changes in disease severity
- Patients should be selected according to predefined clinical criteria
- The primary and secondary outcome measures should be clearly defined
- Allergen extracts and maintenance doses should be defined
- The duration of treatment should be sufficient to allow clinical improvement to occur

Too many studies are planned and initiated using a study design that may not conclusively answer the relevant clinical problem: is allergen-specific immunotherapy a clinically effective treatment? Adherence to the few mentioned criteria will improve the quality of the study and eventually with time close the ongoing debate of the applicability of disease-modifying specific allergy intervention on an evidence-based background.

References

1. Malling HJ, Weeke B. Allergen-specific immunotherapy for allergic diseases.
2. Bousquet J, Lockey J, Malling HJ, et al. Allergen immunotherapy: a practical approach.
3. Malling HJ. Immunotherapy in allergic diseases.
4. Frew AJ, White PJ. Allergen immunotherapy: a practical approach.
5. Altman DG, Bland JM. Measurement error in clinical research.
6. Enas GG, Enas NH. Allergen immunotherapy: a practical approach.
7. Treasure T, MacRae J. Allergen immunotherapy: a practical approach.
8. Devereaux PJ, Mann J, Vandenbroucke JP, et al. Strengthening of reporting of observational studies in epidemiology (STROBE).
9. Bloom BS, de Pauw R. Allergen immunotherapy: a practical approach.
10. Campbell MJ, Julian PR. Allergen immunotherapy: a practical approach.

Discussion

Rak: I think you found a difference between the two groups and that is luck. You cannot control the air, you will not know the severity of the disease if the patients were not enough people.

Malling: You are comparing two groups of patients for two seasons. There is no difference between the two groups.

Cromwell: There is a difference between the two groups but a symptom score is not a marker that it is realistic to use as a marker?

Malling: No. I do not think it is a realistic system. It is not realistic to think that we can control the air in daily practice.

de Weck: It has been shown that immunotherapy improves quality of life.

Malling: I think it is important to evaluate the clinical efficacy of immunotherapy by using symptom and quality of life as very interesting as symptoms. Quality of life is more important.

References

1. Malling HJ, Weeke B. EAACI Immunotherapy position papers. *Allergy* 1993; 48 (Suppl. 14): 9-35.
2. Bousquet J, Lockey RF, Malling HJ (Eds). WHO Position Paper. Allergen Immunotherapy: therapeutic vaccines for allergic diseases. *Allergy* 1998; 53 (Suppl 44): 1-42.
3. Malling HJ. Immunotherapy as an effective tool in allergy treatment. *Allergy* 1998; 53: 461-472.
4. Frew AJ, White PJ, Smith HE. Sublingual immunotherapy. *J Allergy Clin Immunol* 2001; 107: 441-444.
5. Altman DG, Bland JM. How to randomise. *BMJ* 1999; 319: 703-704.
6. Enas GG, Enas NH, Spradlin CT, Wilson MG, Wiltse CG. Baseline comparability in clinical trials. *Drug Inf J* 1990; 24: 541-548.
7. Treasure T, MacRae KD. Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. *BMJ*. 1998; 317: 362-363.
8. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Guyatt GH. In the dark: physician interpretations and expert definitions of blinding in randomized controlled trials. *JAMA*. 2001; 285: 2000-2003.
9. Bloom BS, de Pouvoirville N, Libert S. Classic or Bayesian research design and analysis. Does it make a difference? *Int J Technol Assess Health Care*. 2002; 18: 120-126.
10. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995; 311: 1145-1148.

Discussion

Rak: I think you forgot one major factor which is very important in seasonal studies and that is luck. You have to be lucky because when you do not have enough pollens in the air, you will not be able to find anything. Of course, the outcome is linked to the severity of the disease, but I lost a lot of money and patients during seasons where there were not enough pollens in order to evaluate the data.

Malling: You are completely right. I did not show the results, but we followed these patients for two seasons and the second season was almost without pollen. There was no difference between the groups because they had no symptoms.

Cromwell: There is a reluctance on the part of the regulator authorities to accept anything but a symptom medication score as a basis of evaluating efficacy. Do you believe that it is realistic to use another parameter such as a provocation test as a surrogate marker?

Malling: No. I do not think so. Even the most sophisticated provocation test is an artificial system. It is not comparable with symptoms induced by natural exposure. So I do not think that we could use any substitute for the symptom medication score in clinical studies. In daily practice it is something different.

de Weck: It has become increasingly popular to refine symptom analyses instead of improving quality of life. What do you think about that?

Malling: I think it is very important to include quality of life, but if we are discussing the clinical efficacy, there must be a reduction in disease severity and that means measuring symptom and drug scores. I think that quality of life and other parameters are very interesting as secondary parameters. Maybe in the future we will say that the quality of life is more important than the symptoms, so that we could define this parameter