

Arbeiten aus dem Paul-Ehrlich-Institut
11th International Paul-Ehrlich-Seminar, Bad Homburg 2005
Verlag Chmielorz

Clinical Index-Score and the ALMA Study

Albrecht Bufe, Tim Holland-Letz

Introduction

Many clinical studies in the context of specific immunotherapy lack the application of reliable and accepted criteria to measure the efficacy of treatment. These general criteria are nowadays widely used in all pharmacological studies and are a necessary prerequisite to generate clinical evidence.

The criteria are:

Structural identity of compared groups achieved by randomisation.

Identity of conditions for the observation achieved by blinding.

Precise definition of a primary endpoint in advance and

Exact definition of in- and exclusion criteria.

All these aspects have to be defined in a properly designed study to measure the outcome of specific immunotherapy, which is applied either for the treatment of allergic rhinoconjunctivitis or bronchial asthma. In many studies the precise definition of a primary endpoint is missing.

In the following the development of primary endpoints to be used in the context of allergic studies will be discussed and preliminary results of an international study (ALMA-Study = Allergy Marker Study) will be presented.

Basic requirements for primary endpoints: The validity and reliability.

The two basic requirements for a primary endpoint to be used in a clinical study are the *validity* of the marker and its reliability. Content validity is a factor, which describes whether the method, which is used, actually measures the intended property or if there is some systematic bias. If possible these factors, as judged by comparison with an acknowledged reference criterium (Gold Standard), are usually measured by pearson or spearman correlation coefficient. A correlation close to 1

indicates good agreement. In the absense of a Gold Standard criterium the validity is far from clearly defined. Correlations to other related factors may still be calculated but interpretation of results is unclear, at its best.

If reliable information is available, at least on extreme conditions, for instance in acute asthma, content validity can be determined by measuring the difference between these extreme conditions (discrimination). This factor is called responsiveness and is a prerequisite for validity. It does not, however give any information about the degree of change, which can only be estimated. Responsiveness is generally judged by tests of significance or by the analysis of sensitivity and specificity using binary diagnosis.

In the absense of even this minimal information, judgement of validity can only be performed by medical expertise, for instance by subjectively deciding whether the score components are related to the target property. By default, such an approach will always result in risk of error. The ability of a marker to discriminate between two extreme conditions can be presented by the so-called ROC-analysis. ROC-analysis comprises sensitivity and specificity by measuring the area under the curve of several sensitivities (false positive results) versus several results of specificity (false negative results) under certain cut-off conditions. High sensitivity and specificity is then expressed by a high AUC (area under the curve) of up to 0.99 or 1. Lower sensitivity and specificity is then expressed by AUCs of less than 0.7.

Reliability or repeatability describes the ability of a method to produce nearly identical results when used several times under identical circumstances. Reliability is usually judged by two related coefficients, the *intraclasscorrelationcoefficient* (ICC) and the coefficient of variation (COV). To conduct a reliability study, the method has to be applied under different circumstances. For several time points, the results can then be analyzed by calculating the variance. The ICC is also presented as numbers between 0,0 and 1,0. Good reliability is expressed as a high ICC of up to 1.

Existing studies on clinical index-scores for specific immunotherapy

General marker for a classical study on allergic rhinitis and conjunctivitis are index-scores for clinical symptoms, also called clinical index scores (CIS). In table 1 several studies are shown which have used different scores. The table indicates whether the used scores were validated either by Gold Standard, by discrimination or by retest reliability.

It is shown that only one study by *Marks et al.* uses a score which was correlated to a Gold Standard, namely respiratory function measured by FEV1 and airway hyper-responsiveness. Still the correlation between the marker and the Gold Standard was relatively low. In this study discrimination measured by a method called Crombach alpha was much better and the retest reliability was indicated by up to 0.8.

Juniper et al. evaluated a quality of life score in correlation to rhinitis. When combining the different scores and most single components used in the quality of life

Table 1.

Study	Score	Gold standard	Discrimination	Retest-reliability
Mc Hugh (Allergy 1992)	[comSc: Symp, Self-A, Medi, SPT, NasProv] (Sum rank points = 80)	no	no	no
Bousquet (JACI 1994)	[SF-36 Quality of life score; All. Rhin.] N=227	no	[Mann-Whitney]: 8 out of 9 comp. >0.0001 [Cromb- α]: 0.80	no
Juniper (Clin Exp All 1991)	[LQ: specific for Rhin] n=60	no	[t-test]: significant for sum-score and most single comp.	[ICC]: 0.86
Marks (J Clin Epi 1992)	[comScLq: All Symp & LQ by 58 quest.] n=81	[Corr to FEV1 & AHR]: r=0.16 & 0.38	[Cromb- α]: 0.92	[ICC]: 0.80
Wasserfallen (JACI 1997)	[S]: Skin, Eyes, Nose, Sinuses, Lung (0-5); n=102	no	[Mann-Whitney]: nose, sinuses, lung = 0.0009	0.8 for sum-score
Kilpelainen (Allergy 2001)	[S]: Asthma / Rhinitis, 4 ques- tions (score ?); n=413	no	[Sens]: 0.72-0.99% [Spec]: 0.4%	no
Annesi-Measano (Allergy 2002)	[comSc: Symp, Self-A, MedDia, FamH] (expert opinion) n=269	no	[Sens]: 0.74 [Spec]: 0.83% for SPT+MD	no
Santilli (Ann All Asth Imm 2001)	[comSc: Symp&Medi] (Index + max Med) n=132	no	[t-test?]: significant [Cromb- α]: approx. 0.85	no
Bufo (Allergy 2004)	[comSc: Symp&Medi] (2 index + medSc) n=133	no	[Mann-Whitney]: Clin Index Score >0.0001	0.89 for CIS

questionnaire retest reliability was relatively high (ICC=0.86). Furthermore, *Wasserfallen et al.* demonstrated a good discrimination for his clinical score with an acceptable retest reliability of ICC=0.8. Also, *Bufo et al.* used a combined clinical index score adding clinical symptoms and medication. By retrospective analysis the authors found an acceptable discrimination for their clinical index score and a relatively high retest reliability of 0.89 for ICC.

Still, taking these results together no general recombination can be given for acceptable markers measuring severity of allergic disease and there are only very few studies really showing content validity and reliability of the markers used.

ALMA study design

These observations were the reason for the assembly of an international study group (ALMA Study Group=Allergy Marker Study Group), getting together the Experimental Pneumology and the Department of Medical Informatics Biostatistics and Epidemiology from the Ruhr-University Bochum, Germany, responsible for design, coordination, and analysis of the study, the Fraunhofer Institute Hannover responsible for the performance of the study, especially in their pollen chamber, the Sanquin Institute in Amsterdam responsible for measuring serology and antibodies and finally the University of Amsterdam (AMC) responsible for the analysis of cellular immunology markers.

The aim of the ALMA study was to evaluate known and accepted markers for severity of allergic symptoms and the identification and evaluation of early markers for allergic symptoms indicating efficacy of specific immunotherapy. Figure 1 shows

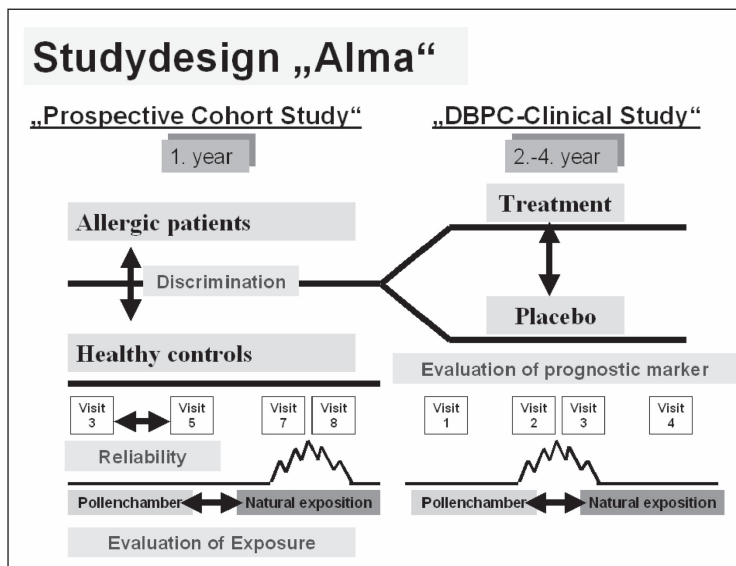


Fig. 1. Study design of the ALMA I and ALMA II study. Preliminary results of the ALMA I study are presented in this article, ALMA II study is in the process of planning.

the study design of the ALMA study as a prospective cohort study for one year, comparing a group of grass pollen allergic patients (n=60) with healthy controls (n=60). These patients were twice analyzed for several markers; (see below) outside the pollen season, just before and during a provocation in the pollen chamber, as well as twice during the following pollen season. All data underwent a discrimination analysis (ROC-analysis) between the patient and a healthy control groups. Reliability/ Repeatability of the different markers, between the respective two visits, were measured by ICC. By correlation the association between pollen chamber and pollen season results were compared. Once certain reliable and valid markers will be identified, they will be evaluated in a double blind placebo controlled clinical study to identify early prognostic markers for the outcome of the clinical study.

The markers measured in the ALMA study were:

Symptom-questionnaires for nasal, eye, and bronchial symptoms, quality of life scores and function and provocation tests such as lung function, nasal flow measurements, nasal provocation, and NO-exhalation either bronchial or nasal.

Serological and cellular markers such as total IgE and specific IgE towards grass pollen allergens as well as IgG, IgG4, or IgA.

Cellular markers were T-cells in number and T-cell activation in cytokine release of these T-cells basically measured by ELISPOT analysis.

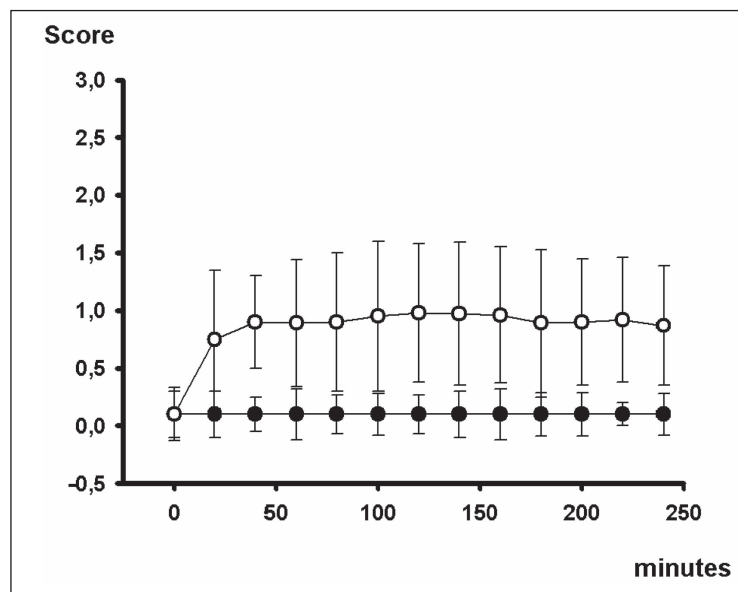
Preliminary results of ALMA study

The results presented here are only preliminary.

Figure 2 shows results for one clinical symptom, namely nasal itching. The discrimination of this symptom, between the two study groups at the first provocation visit in the pollen chamber over 240 minutes, demonstrates that the symptom maximum is reached relatively early after 20 minutes and remains constant in the diseased study group. With an acceptable variation this symptom is highly discriminative under pollen chamber provocation conditions.

For further presentation a “sum symptom score” consisting of nasal, eye, and airway symptoms was measured calculated. Patients with pollen allergic rhinitis show a significant increase of symptoms as compared to those after provocation and to the healthy control. The “sum symptom score” measured in the pollen chamber is comparable to those symptom measurements during natural exposition especially when pollen counts are significantly increased. Reduced pollen counts are associated with a lower “sum symptom score”.

Fig. 2. Results of measurement of symptom “nasal itching” in the pollen provocation chamber. Score indicates “values given for the specific symptom”, minutes indicates “time of provocation in the pollen chamber”.



ROC analysis of these symptom scores reveals a high sensitivity and specificity when used in the pollen chamber. Both sensitivity and specificity are lower when applied during the pollen season. Related to that retest-reliability is high when the assays are used in the pollen chamber (ICC = 0,97), but relatively low when used in the pollen season (ICC = 0,7).

The result of the latter can be explained by the pollen exposition, which was clearly different at the two time points of measurement. A high correlation between the “sum symptom score” including nasal, eye, and lung symptoms and the “total nasal symptom score” (TNSS) as recommended by the FDA of 0.95 as long as applied during pollen chamber measurements. The correlation between pollen chamber and natural exposition measurements is 0.74 and thus acceptable.

Interestingly the TNSS has a higher sensitivity and specificity compared to the measurement of the single symptoms, which are content of the TNSS. This indicates that the combination of several symptoms is better than the measure of just one item.

The quality of life score, according to Juniper *et al.*, is best used during natural exposition and senseless when used in the pollen chamber since quality of life does not change in such a short period of time such as two hours. This score is highly sensitive and specific in our hands and shows a retest-reliability of 0.82 ICC which is comparable to what Juniper has published.

Nasal flow measurements are not very sensitive and specific. The same is true for NO-exhalation although the latter has a very good retest-reliability (ICC=0.9). The measurement of serum IgE antibodies towards recombinant grass pollen allergens is highly specific and sensitive for the diagnosis of allergy and shows a very good retest reliability.

There is a correlation between the antibody titer towards recombinant grass pollen allergen Phl p 1 with the “sum symptom score”. Taking a closer view on the data, it reveals that this correlation is only relevant when the titer is above a certain level, of at least 40 U/L. We found no correlation between IgG4 and IgE with the “sum symptome score”.

Summary of preliminary results of the ALMA study

Pollen chamber provocation with grass pollen could be a Gold Standard for “sum symptome scores” measuring symptoms of allergic rhinitis. Measuring “sum symptome scores” such as TNSS is more sensitiv and reliable than measuring single symptoms. Nasal flow and NO-exhalation are less sensitive and less specific markers for allergic rhinitis.

IgE towards recombinant grass pollen allergens is highly sensitive for diagnosis of allergic rhinitis and more so, than IgE towards only grass pollen extract. IgE towards recombinant Phl p1, 2, 4, and 5 correlates with symptoms as measured in pollen chamber provocation. IgG4/IgE ratio only weakly correlates with symptom scores.

Acknowledgement

This work was designed and performed by the ALMA study group: M. Larbig, J. Hohlfeld, A. Schauerte, N. Krug (all Fraunhofer Institute Hannover); E. Wierenga, M. Kapsenberg (all AMC, University of Amsterdam); R. van Ree (Sanquin, Amsterdam); T. Holland-Letz, A. Hüsing, H.-J. Trampisch (all Med. Informatics, Biomathematics and Epidemiology, University Bochum); K. Gehlhar, A. Bufe (all Experimental Pneumology, University Bochum). It was sponsored by HAL Allergy, Haarlem, Netherlands.

Discussion

Bergmann: A question regarding the quality of life following the stay in the pollen chamber: I guess the question in the questionnaire of Elisabeth Juniper is "How did you feel in the last week?". So if you only stayed in the pollen chamber for hours then you cannot expect that patients answered what had been in the last 7 days. Is that right?

Bufe: But there are also symptoms measured in there and it is interesting that the symptom score does not really measure the difference.

Bergmann: May I congratulate you on the NO measurements, which is very good. But you have to keep in mind that NO levels of course are influenced, both by allergic rhinitis and also by asthma, that is by an eosinophil influx into the bronchi.

Falagiani: For another bronchial provocation test in the pollen chamber you mentioned that usually just the early-phase reaction was checked, I think also because of organizational problems. It is not difficult to keep the patient under control for 6 - 8 hours after the challenge to also check their late-phase reaction. Nevertheless, I believe that from the clinical point of view the late-phase reaction is most important and even most troublesome for the patient. What's your opinion about this point?

Bufe: For me as a clinician this definitely is an important point. But as you have seen we only measured that for NO, for some other clinical markers. But since these patients have been selected on the inclusion criterion of allergic rhinitis as a major diagnosis, we do not see any changes in lung function or any other things which occur later. But we have not yet measured anything for late-phase reaction. Maybe there will be data coming up by the T cell reactivity which I cannot tell you yet.

André: Just a question about this clinical score. Do you use the classical scoring system everyone is using: zero: absent, one: present but non disturbing, two: present and disturbing, three: present and very disturbing? That score is very subjective.

Bufe: Of course it is subjective, but I think we already had this discussion this morning. The most interesting result of this study is, in my opinion, that the most subjective parameter is the most sensitive one. You just have to think about that, but this brings the clinician again back to the stadium. I hope we will find other markers

166

which indicate a change which can tell us “okay, this is the right patient or the wrong one”. In this context, we have seen some other data this afternoon. I do not know whether I answered your question correctly, but it is clearly subjective and obviously it is the best.

André: Another short question: Nasal flow perfectly applies to pollen stimulation. How does it apply to a chronic situation, like rhinitis due to mites where patients are blocked all the year?

BuFe: We have not tested that, because in this study we excluded those patients beforehand. Otherwise we could not have tested the whole test system.