A critical appraisal of analyzing nasal provocation test results in allergen immunotherapy trials*

Nicole Graf[1,2], Beni Dinkel[1], Horst Rose[3], Ludwig A. Hothorn[4], Daniel Gerhard[4], Pål Johansen[5], Thomas M. Kündig[5], Ludger Klimek[6], Gabriela Senti[1]

[1] Center for Clinical Research, University Hospital Zurich, Zurich, Switzerland

[2] graf biostatistics, Winterthur, Switzerland

[3] Rose Pharma-Consulting, Burgdorf, Germany

[4] Institute of Biostatistics, Leibniz University of Hannover, Hannover, Germany

[5] Department of Dermatology, University Hospital Zurich, Zurich, Switzerland

[6] Center for Rhinology and Allergology, Wiesbaden, Germany

Rhinology 52: 0-0, 2014

DOI:10.4193/Rhino13.145

***Received for publication:**

September 12, 2013

Accepted: November 6, 2013

Abstract

Background: The statistical analysis of nasal provocation tests is very complex. We compared the conventional analysis with the maximally selected test statistics and the hierarchical ordered logistic model.

Methods: We re-analyzed data from a trial with 112 patients suffering from grass pollen allergy. The patients had been randomized to receive either intralymphatic immunotherapy (ILIT) or subcutaneous immunotherapy (SCIT).

Results: The conventional analysis indicated that the logarithmized ratio between the pre- and the post-treatment threshold concentration was significantly lower for ILIT than for SCIT. The maximally selected test statistics was used to test different threshold symptom scores that would imply positive clinical symptoms at the given allergen concentration. A threshold score of 3 maximised the difference in improvement between the ILIT and the SCIT groups. The hierarchical ordered logistic model does not take threshold allergen concentrations as the basis for analysis, but the single scores measured at each concentration. This approach simultaneously considers the treatment effect (ILIT versus SCIT), the time effect (pre- versus post-treatment), and the dose effect (different allergen concentrations). The hierarchical ordered logistic model revealed that the clinical improvement was greater after ILIT than after SCIT.

Conclusion: As the choice of method can affect the outcome, guidelines for analysis are highly needed.

Key words: nasal provocation test, immunologic desensitization, statistical data analysis

Introduction

Kirkman in 1835 and Blackley in 1873 were the first to experimentally reproduce symptoms of allergic rhinitis in sensitized individuals by applying pollen to the nasal mucosa. Today, the direct nasal allergen challenge model is used to evaluate the response of the nasal mucosa in allergic rhinitis for diagnostic reasons and to evaluate different treatments [(1)]. This nasal challenge test (NCT) or nasal provocation test (NPT) uses either a single provocation (supra-threshold) or a series of successive provocations with increasing allergen doses separated by at least 10 min intervals (titrated provocation) [(1,2)]. The first method is performed with relatively low effort within 30 to 45 min, however, it allows only a qualitative evaluation and is thus mainly used for diagnostic purposes. The titrated nasal provocation

(tNPT) gives more quantitative information on the sensitivity of the mucosa and thus allows for evaluations of different treatment options in a pre- / post manner. However, this test can take up to 180 min [1]. The target organ is challenged with titrated allergen doses, starting with the lowest dose, after which subjective symptoms are being scored by patient and clinician, and objective measures are taken simultaneously [1]. Allergen immunotherapy trials especially require objective methods for efficacy evaluation, since effects of varying allergen exposure and intake of rescue medication have tremendous influence on the reported symptoms [3]. Guidelines by the World Allergy Organization task-force recommend the use of provocation tests as secondary outcomes in immunotherapy studies [3,4], and the European Medicines Agency guideline on the clinical development of products for immunotherapy suggests that provocation tests may be used as primary endpoints in dose-finding studies [5]. While NPT is usually regarded to be a reliable and valid method, there are no internationally standardized procedures for NPT. The applied methods vary with respect to the factor of allergen dilution, the time point for assessment of the symptoms, the clinical symptoms and objective parameters to be assessed, as well as how symptoms are to be assessed. While these methodical shortcomings are well-known and have been responded to by various authorities and researchers [1,2,6], the problems connected to the complexity of the statistical analysis in tNPT have remained unmentioned and probably largely unrealised. We used data from a randomized controlled hay-fever immunotherapy trial [7] to highlight some problems of the statistical analysis of NPT data and discuss the validity of various statistical methods.

Materials and methods

In the monocentric open-label trial at hand, 165 patients with grass pollen-induced rhinoconjunctivitis were asymmetrically randomized (3:2) to receive either three intralymphatic injections over two months or conventional immunotherapy, i.e., 54 subcutaneous (s.c.) injections with pollen extract over three years [7]. Of the 99 patients randomized to the s.c. immunotherapy (SCIT) group, 54 started the treatment. Out of 66 patients randomized to the intralymphatic immunotherapy (ILIT) group, 58 started the treatment. The efficacy of SCIT and ILIT was tested and compared based on tNPT data at baseline and after 4 months, 1 year, and 3 years. In the current statistical investigation, only the data for baseline and 4 months were utilized. tNPT was performed according to standard procedures [1]. Patients were challenged with four increasing concentration levels of the allergen, 10^2, 10^3, 10^4 and 10^5 SQ-U/ml grass pollen extract. A symptom sum score ranging from 0 – 6 points was recorded with three groups of symptoms, namely nasal secretion, sneezing and remote symptoms, each scored at 0, 1 and 2 as described in Table 1. The lowest pollen concentration inducing a total score of 4 or higher (out of 6) was defined as the maximal tolerated pollen concentration. For reasons of safety, the pollen dose was therefore not further escalated for these patients.

Table 1. Symptom score assessment in the nasal provocation test (NPT).

	Score		
Symptom	**0**	**1**	**2**
Nasal secretion	none	mild	severe
Sneezing	0-2	3-5	>5
Remote symptoms	none	Lacrimation or pruritus of the ear/palate	Conjunctivitis, chemosis, urticarial coughing or shortness of breath

Results and Discussion

Conventional statistical approach

A common statistical approach in the evaluation of allergen-specific immunotherapy consists in calculating the ratio between pre- and post-treatment threshold concentrations and comparing the calculated ratios between the treatment groups, e.g. by a non-parametric Mann-Whitney U test [8]. The advantages of such an approach are that the statistical analysis is simple and it reflects the clinical understanding that an improvement of allergy symptoms is relevant only when at least a tenfold increase in the tolerated allergen provocation dose is obtained. The results from the exact Mann-Whitney U test indicate that the ratio of the pre-treatment/post-treatment threshold concentrations is significantly lower ($p < 0.001$) for ILIT than for SCIT (Figure 1).

A major disadvantage of this type of analysis is that it considers only threshold allergen test doses and partly ignores valuable information of the collected data, i.e., the sum of scores measured at the different allergen test concentrations. Another disadvantage is that the resulting p-value depends directly on the pre-defined threshold symptom score. Since the threshold symptom score defines the allergen concentration at which a patient is assumed to react positively, it has a direct impact on the endpoint, which in turn influences the results of the statistical analysis. Finally, it is difficult to define the threshold score before having performed the provocation test. An optimal threshold score would discriminate between the patients from different treatment groups. However, if a too high threshold score is chosen, the majority of all patients may possibly react only at the highest allergen concentration for the test at the baseline time point. As a consequence, the only improvement

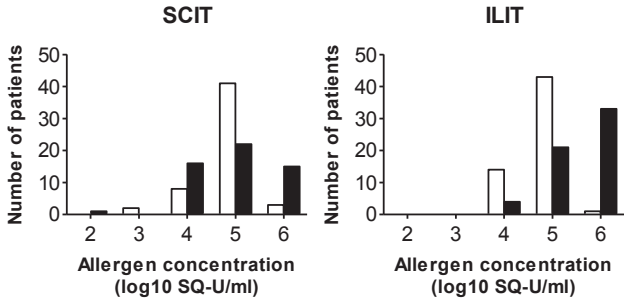

Figure 1. Number of patients reaching the reaction threshold at allergen provocation concentrations 10^2, 10^3, 10^4 or 10^5 SQ-U/ml grass pollen extract before (open bars) and after therapy (closed bars). For patients not reacting at 10^5 SQ-U grass pollen extract, the threshold was set at the next higher allergen concentration, i.e., 10^6 SQ-U/ml. A comparison of the ratio of pre-treatment/post-treatment threshold concentrations indicates that the ratio is significantly lower for ILIT than for SCIT ($p < 0.001$).

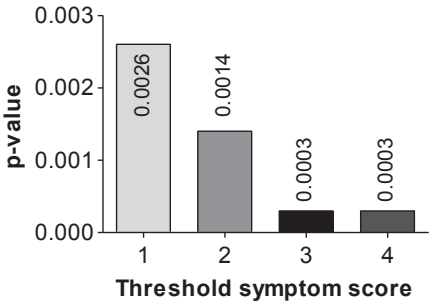

Figure 2. P-values calculated for the comparison of improvement in nasal allergen tolerance at different threshold symptom scores. Threshold concentrations at baseline and after four months of immunotherapy were compared, and a min-P approach for threshold scores 1, 2, 3, and 4 was applied [9].

possible is that these patients will not reach the threshold score at the highest concentration after the treatment. In such a situation, the test will probably not be sensitive enough to reveal therapeutic efficacy or significant differences between patients treated by different methods.

Maximally selected test statistics

The method of maximally selected test statistics can deal at least with the two latter problems by testing all possible threshold scores and afterwards selecting the best threshold score with the smallest p-value. As more than one hypothesis is being tested, the problem of an inflated chance for a type I error, i.e., falsely rejecting the null hypothesis, has to be taken into account. The method of maximally selected test statistics precisely accounts for this problem as it makes an adjustment against the multiplicity of using several threshold scores [9]. However, if the threshold symptom score is set at a rather low level, the threshold concentration for higher scores will as a consequence be unknown. Thus, the maximally selected test statistics is ideal for situations where a rather high threshold score is chosen. Figure 2 illustrates that 3 represent an optimal threshold score to demonstrate an improvement in nasal tolerance when comparing baseline NPT data with data obtained four months after starting immunotherapy. A threshold score of 4 resulted in a comparably good, though marginally higher p-value. Based on a threshold score of 3, it can be shown that the ratio between the pre- and post-treatment threshold concentrations is significantly different between the ILIT and SCIT treatment groups ($p < 0.001$). If another threshold score is chosen, the p value will increase, hence, the significance of the results will be reduced. Thus, while the conventional statistical approach using the Mann-Whitney U test resulted in the same conclusion, it was not based on the optimal threshold score of 3.

Hierarchical ordered logistic model

An alternative statistical approach for analysis of the same set of tNPT data is based on an integration of all data from all tested concentrations and all time points. Here, the time effect (before and after immunotherapy), the treatment effect (the sum of total symptom scores for the two treatment groups), and the dose effect (four different allergen test concentrations) can be analyzed using a factorial design, in which the single main effects and the interactions between factors are calculated. The interaction between the treatment effect and the time effect is of special interest as it describes the change in response due to the effect of the treatment. It may be discussed if any further interactions need to be modelled, i.e., a dose-treatment interaction. Our hierarchical ordered logistic model is restricted to incorporate the main effects and the treatment-date interaction (Table 2). As the scores represent ordered categorical data, the use of an ordered logistic model is proposed [10,11]. The logistic model estimates from Table 2 can be transformed into the odds ratio scale by taking the exponent. In the current clinical example, these estimates represent the chance of reaching lower score categories with one therapy (ILIT) than with the other therapy (SCIT) considering the variability in each patient's symptoms by modelling the between- and within-subject effects by estimating additional variance components in a mixed model framework. This approach supports the interpretation that after 4 months, the improvement in nasal allergen tolerance after ILIT was significantly greater than after SCIT ($p < 0.001$). The exponent of 1.797 suggests that compared to SCIT, ILIT has a six times higher

Table 2. Test of interaction terms based on hierarchical ordered logistic model parameter estimates.

Time	Treatment	Estimate	Std. Error	p-value
before vs. after treatment	SCIT vs. ILIT	1.797	0.270	< 0.001

chance of reaching lower scores after the treatment.

The nature of tNPT data, which are used in the context of testing efficacy of allergen immunotherapy, is highly complex and requires adequate statistical handling. Apart from the conventional approach, which compares allergen concentrations being reached at predefined threshold scores, alternative strategies are worth being discussed especially to be able to integrate the whole body of harvested data. Such data typically include varying threshold values for the provocation test, varying test concentrations of the allergen, as well as varying time points for testing. It seems, however, that there is no best method, i.e., a flaw-less method that could be generally recommended in clinical studies evaluating the efficacy of allergen immunotherapy. Firstly, all utilized statistical approaches work with sum of symptom scores. Calculating a sum assumes comparable scales in the three categories nasal secretion, sneezing, and remote systems and further implies that the symptom scores can be summed. An increase in the sneezing score from 0 to 1 for example is assumed to be equivalent in weight to an increase in the nasal secretion score from 1 to 2. Secondly, a potential correlation between the categories is not accounted for, i.e., categories being possibly correlated each have the same weight as a single uncorrelated category. These questions typically arise when additive rating scales have to be validated, and should therefore also be addressed with respect to tNPT.

The conventional statistical approach of analyzing tNPT data is based on a predefined threshold symptom score. The problematic necessity of predefining a threshold score can be overcome by a maximally selected test statistics approach, thus, choosing the best threshold score after having performed the allergen provocation test and obtained the data. However, and similar to the conventional approach of analyzing tNPT data, the method of maximally selected test statistics considers only threshold concentrations and ignores the large quantity of potentially valuable information represented by the score values measured at each concentration. Moreover, the maximally selected test statistics has limitations in situations where the provocation test is discontinued as soon as a low score was reached.
The hierarchical ordered logistic model is based on the symptom scores measured at each allergen concentration, but has limitations with regards to potentially missing values. A statistical analysis that takes into consideration score values instead of allergen concentrations necessarily has to deal with missing values as tNPT´s are usually stopped for ethical reasons as soon as patients reach a certain score value. These missing values are informative, since they reflect a very high test reactivity. As a consequence, patients with higher test reactivities are lost, and this may violate major interests of the clinical trial. Moreover, missing score values may lead to an unbalanced design. Nonetheless, this sort of data analysis may be interesting and

Table 3. Advantages and disadvantes of different methods of analysis.

Method of analysis	Advantages	Disadvantages	Recommendation
Conventional approach	Statistical methods are simple and interpretation of results is straight forward. The results reflect the clinical understanding that improvements are measured on a tenfold scale.	The threshold symptom score has to be defined before the start of the study. Valuable information, i.e., the scores measured at each allergen concentration, is ignored in the statistical analysis.	To be used if threshold symptom score can be defined reliably before start of study and if provocation is discontinued at a rather low threshold symptom score.
Maximally selected test statistics	The best threshold symptom score is chosen ex post. The results reflect the understanding that improvements are measured on a tenfold scale.	Valuable information, i.e., the scores measured at each allergen concentration, is ignored in the statistical analysis.	To be used if definition of threshold symptom score is uncertain and if provocation is discontinued at a rather high threshold symptom score.
Hierarchical ordered logistic model	All gathered information is used in the statistical analysis. This analysis may thus prove more sensitive to possibly smaller differences.	Statistical methods are complex. The analysis has to deal with the problem of missing values.	To be used if missing values pose no major problem and if help is provided by a statistician who can carry out the analysis.

advantageous provided that the problem of missing values can be overcome, e.g. by using lower concentration in the allergen provocation tests combined with a lower dilution factor and higher threshold scores or by imputing missing values.
In conclusion, both analyses of threshold symptom scores and of symptom score values have limitations with respect to the evaluation of allergen provocation tests before and after allergen-specific immunotherapy. However, this study demonstrates that comparable statistical results can be obtained with different statistical approaches. Depending on the characteristics of the population and the details of the provocation test, one statistical approach may prove more useful than another. Table 3 summarizes advantages and disadvantages of the presented methods of analysis. It may also be sensible to plan more than one statistical analysis, either as an attempt to perform a sensitivity analysis or as an option to choose the most sensitive analysis. Such an approach must of course also take into consideration the multiplicity problem. It would be highly desirable firstly that rating scales of tNPT would be properly validated, secondly that international guidelines would be formulated with respect to the procedure of NPT, and thirdly that a recommendation can be made with respect to the statistical approach of analysing tNPT data.

Authorship contribution

NG designed the study, analysed data and wrote the manuscript.
BD collected data.
HR, LAH and DG analysed data and revised the manuscript.
PJ, TMK and LK concepted, wrote and revised the manuscript.
GS designed the study and wrote the manuscript.

Acknowlegdement

None

Conflicts of Interest

The authors disclose no conflict of interest.

References

1. Riechelmann H, Bachert C, Goldschmidt O, et al. [Application of the nasal provocation test on diseases of the upper airways. Position paper of the German Society for Allergology and Clinical Immunology (ENT Section) in cooperation with the Working Team for Clinical Immunology]. Laryngorhinootologie 2003; 82: 183-188.
2. Hellings PW, Scadding G, Alobid I, et al. Executive summary of European Task Force document on diagnostic tools in rhinology. Rhinology 2012; 50: 339-352.
3. Klimek L, Pfaar O. A comparison of immunotherapy delivery methods for allergen immunotherapy. Expert Rev Clin Immunol. 2013; 9: 465-474.
4. Canonica GW, Baena-Cagnani CE, Bousquet J, et al. Recommendations for standardization of clinical trials with Allergen Specific Immunotherapy for respiratory allergy. A statement of a World Allergy Organization (WAO) taskforce. Allergy. 2007; 62: 317-324.
5. Guideline on the clinical development of products for specific immunotherapy for the treatment of allergenic diseases. European Medicine Agency document: CHPMP/EWP/18504/2006.
6. Mortemousque B, Fauquert JL, Chiambaretta F, et al. [Conjunctival provocation test: recommendations]. J Fr Ophtalmol. 2006; 29: 837-846.
7. Senti G, Prinz Vavricka BM, Erdmann I, et al. Intralymphatic allergen administration renders specific immunotherapy faster and safer: a randomized controlled trial. Proc Natl Acad Sci U S A. 2008; 105: 17908-17912.
8. Senti G, Graf N, Haug S, et al. Epicutaneous allergen administration as a novel method of allergen-specific immunotherapy. J Allergy Clin Immunol. 2009; 124: 997-1002.
9. Pollard KS, Gilbert HN, Ge Y, Taylor S, Dudoit S. multtest: Resampling-based multiple hypothesis testing. R package version 2.14.0.
10. Christensen RHB. Ordinal: Regression Models for Ordinal Data. R package version 2012.09-11.
11. Agresti A. Categorical Data Analysis. Second edition. New York: Wiley, 2002.

Gabriela Senti
Center for Clinical Research
University Hospital Zurich
8091 Zurich
Switzerland

Tel: +41-44-634-5509
Fax: +41-44-634-5505
E-mail: gabriela.senti@usz.ch