REVIEW ARTICLE

Methodological Issues in the Diagnostic Work-up of Food Allergy: A Real Challenge

M Gellerstedt,[1] U Bengtsson,[2] B Niggemann[3]

[1] Department of Informatics, University West, Trollhättan, Sweden
[2] The Asthma and Allergy Research Group, Department of Respiratory Medicine and Allergy, Sahlgrenska University Hospital, Gothenburg, Sweden
[3] Department of Pneumology and Immunology, University Children's Hospital Charité, Berlin, Germany

■ **Abstract**

The standard of reporting in diagnostic studies has generally been low. Fortunately, this issue has begun to be addressed in recent years through the discussion of important methodological issues in educational series, textbooks, and checklists. Double-blind, placebo-controlled, oral food challenges (DBPCFC) are considered to be the gold standard for diagnosis of food allergy. However, there is no consensus regarding how to interpret the outcome and how to define positive and negative provocations in DBPCFC. Furthermore, since most theories on the diagnosis of food allergy rely on the assumption that the DBPCFC has a high accuracy, this accuracy must be formally statistically evaluated.

In this review, we discuss essential methodological issues for diagnostic accuracy studies in general and for oral food challenges in particular and discuss the importance of methodological issues as a guide for forthcoming studies of diagnostic procedures.

Key words: Diagnostic accuracy. Food allergy. Food challenge. Methodological issues.

■ **Resumen**

La información habitual en estudios diagnósticos es en general poco precisa. Afortunadamente, este asunto ha empezado a tratarse en los últimos años a través del debate sobre importantes cuestiones metodológicas en colecciones educativas, libros de texto y listas de verificación. La prueba de provocación doble ciego controlada con placebo (DBPCFC, por sus siglas inglesas) se considera la prueba de referencia para el diagnóstico de las alergias alimentarias. No obstante, no existe consenso en cuanto a la interpretación de los resultados y en cómo definir las provocaciones positivas y negativas en el DBPCFC. Lo que es más, dado que muchas teorías sobre el diagnóstico de la alergia alimentaria parten de la asunción de que el DBPCFC es de una gran precisión, aseveración que deberá ser evaluada estadísticamente.

En esta reseña, tratamos asuntos metodológicos esenciales para los estudios de la precisión diagnóstica en general y para las provocaciones alimentarias orales en particular. También nos referimos a la importancia de las cuestiones metodológicas como guía para los futuros estudios sobre los procedimientos diagnósticos.

Palabras clave: Precisión diagnóstica. Alergia alimentaria. Provocación con alimentos. Metodología.

Introduction

In order to give the best possible care, a correct diagnosis is fundamental. Naturally, a new diagnostic test must be carefully evaluated prior to its use in clinical practice. However, compared to the rigorous requirements before introducing a new drug there are no similar requirements for adopting a new diagnostic test. This is remarkable, since a diagnostic test in itself could be regarded as a kind of intervention. In the case of food allergy, diagnostic shortfalls could actually harm the patient by leading to implementation of unjustified elimination diets and treatments or by missing effective therapeutic diets. Furthermore, inaccurate diagnoses can generate unnecessary costs.

In clinical trials, methodological issues such as double-blind techniques, the use of controls, and randomization are well-established concepts, and comprehensive textbooks and guidelines have been available for a number of years. The corresponding theoretical development regarding evaluation of new diagnostic tests currently lags behind that of therapeutic studies, and consequently, the reporting of such studies is often of poor quality [1]. Encouragingly, since the early

1990s diagnostic theory has undergone a renaissance and methodological issues have been highlighted in a number of articles and educational series [2-5], as well as in textbooks [6-7]. Proposals have also been made for the development of frameworks [8-10] and a checklist for reporting studies of diagnostic accuracy [11]. Furthermore, since the theory regarding the development of clinical reference values partly overlaps with the theory of diagnostic accuracy, some good advice can also found in guidelines [12-17]. In the declaration of Helsinki it is pointed out that "The primary purpose of medical research involving human subjects is to improve prophylactic, diagnostic and therapeutic procedures and the understanding of the aetiology and pathogenesis of disease" [18]. Naturally, it is important that a high scientific quality is maintained in the reporting of possible improvements in diagnostic procedures.

The double-blind, placebo-controlled food challenge (DBPCFC) is regarded as the gold standard for diagnosis of food allergy. In 1988, the first manual on how to perform a DBPCFC was published [19] and since then the technique has been used in practice, refined, and improved [20-23]. A recently published position paper includes a summary of methodological issues and gives advice on how to perform DBPCFC [24].

When a new potential test is evaluated in a test accuracy study, the results from the new test are compared to the results from the reference test. It is worth mentioning that DBPCFC is used as the gold standard in 2 different situations: in the diagnostic work up of an individual patient [24] and as the reference test in studies where a new test is being evaluated. There are several potential pitfalls in the diagnostic work up for food allergy, [25]. Furthermore, the DBPCFC procedure includes risks and is time consuming for both the patient and physician. Therefore, it is desirable to further improve the diagnostic work-up of these patients and to determine whether the DBPCFC could be replaced by alternative tests in this situation [26-33]. In this article, we will focus on the evaluations of new potential tests to replace DBPCFC and discuss methodological issues in the context of food allergy, highlighting some of the most important issues for test accuracy studies.

Settings: Selection Bias, Spectrum Effect, and Design

When the discriminatory power of a new test is evaluated (in terms of sensitivity and specificity), a case–control study is a common design. In such a design, the choice of cases and controls is crucial. A selection bias may be introduced and the results found in the study may not be valid in clinical situations. For instance, if patients with proven food allergy are compared with truly healthy control subjects in a study, the diagnostic test may show good discriminatory power between these 2 distinct groups. When the test is applied in clinical practice on patients with suspected food-related symptoms, you may be faced with either true food allergy or with food aversion or subjective symptoms. Thus, when the test is used in clinical practice on groups of patients with less-evident differences, the discriminatory ability of the test may not be equally strong. We therefore recommend using a cohort design as the default option, where patients who are suspected to have food allergy are recruited and tested with the new test and the reference test independently [34]. To be able to compare different studies it is also important that inclusion and exclusion criteria are identical.

It is also quite well established that test accuracy in terms of sensitivity and specificity may vary between settings; this is commonly referred to as spectrum bias [35-40] or the spectrum effect [41]. Moreover, it must be taken into account that if the population consists of subpopulations with different characteristics possibly related to the test, the accuracy may also vary within a given setting [42]. There are several important potential differences between settings affecting diagnostic features of food allergy [24]. These include age, presence of concomitant disease such as atopic eczema, the criteria used for determining positive oral challenge (especially regarding subjective symptoms), the length of the observation period during oral challenges, whether or not the patient was on an elimination diet before challenge, patient compliance with diets, and whether or not an open challenge preceded the DBPCFC. Therefore, it is important to carefully describe the sample of patients and study conditions in order to facilitate comparisons between studies. Given the possibility of differences between subpopulations even within the same study population, it is recommended that possible differences in diagnostic accuracy be evaluated between such subpopulations.

Standardization of Measurement and Interpretation

The sources of variability in more objective measures such as skin prick test (SPT), specific serum immunoglobulin (Ig) E, or the atopy patch test are primarily biological variation between and within patients and analytical variation [43]. Analytical variation consists of both random variation and possible systematic variation. Possible sources of systematic variation include different laboratories, instruments, and observers. If the same observer uses the same instrument and measures the same specimen, repeated measurements will vary—ie, there will still be random variation. There may also be variation due to the way in which the collected specimen is handled before it enters the analytical process; this is referred to as the preanalytical variation.

For diagnostic purposes it is important to study the different sources of variability. A standardization of the measurement process to minimize the analytical variation is desirable since it increases the discriminatory power of the diagnostic test. It is rather common to assume that objective measures have high interobserver and intraobserver reliability. Since this is not self-evident, it should be investigated. For instance, a variable such as SPT has high variation both between and within individuals [44], whereas a variable such as body height in adults may vary between individuals but is constant within an individual. In the case of food allergy, one can use objective continuous variables such as specific serum IgE, but the presence or absence of clinical reactions (eg, urticaria or flushing) is usually required for diagnostic purposes. In this situation, a low within-patient variability means that the symptom should occur after each active provocation and be absent after a placebo provocation. If this is the case, an oral challenge procedure using 1 active and 1 placebo provocation is regarded as sufficient. In DBPCFC, the variability in preparation of the provocation meal is a source of preanalytical variability. The presence or absence of nonobservable (subjective) signs such as nausea, gastrointestinal pain, or burning of the tongue is more difficult to relate to the offending food. This can have 2 explanations: either the clinician cannot confirm the symptoms or the symptoms could also occur for other reasons.

There are also potential differences between observers regarding interpretation of subjective symptoms (even if it is observable), such as whether flushing is judged on equal terms and what happens if the patient also presents mild flushing on placebo. In the latter situation, for instance, does that mild flush constitute a threshold that must be significantly exceeded on active provocation or should the challenge be regarded as a failure or even negative? It is important to develop precise standards for the collection and interpretation of these "soft" measurements in order to achieve the highest reproducibility possible. This is necessary in order to judge whether the test is positive or negative, and it represents important information when different studies are compared. It has been demonstrated that a systematic approach to interpreting symptoms can give high interobserver reliability, even in cases with vague symptoms like gastrointestinal pain [45]. To confirm whether or not subjective reactions are associated with the food tested, several provocations must be used—for instance, 3 active and 3 placebo provocations. This is necessary since the proportion of positive placebo provocations may be as high as 35% [46].

Reference Tests

Since new diagnostic tests are compared with a reference test it is desirable that the reference test is 100% certain. However, such an error-free, true gold standard is uncommon. Thus, in most situations the reference test could actually be erroneous, meaning that we have to use an imperfect reference test and take this into account [47]. Naturally, it is desirable to use a reference test with the highest accuracy possible in order to make evaluations of a new test meaningful.

DBPCFC is regarded as the gold standard for diagnosis of food allergy. It is also used as a reference test in studies where a new test is investigated. Consequently, we cannot evaluate the reference test against the gold standard since these are one and the same. Furthermore, the accuracy of DBPCFC depends on how it is conducted and how the results are interpreted, as discussed above. Thus, for diagnosing food allergy, the test considered to be the gold standard is not evidence based and its accuracy is unknown. This is frustrating since estimates of the prevalence of allergy to different kinds of food depend upon the accuracy of the test. For instance, in a study of food allergy in adults with subjective symptoms, there were around 35% positive DBPCFCs [45]. Does this mean that the prevalence in those patients was 35%? Since a part of the positive tests may have been false positives, and since there could also be false negatives, we cannot tell. Is it possible that all positive tests are false—ie, that food allergy with subjective symptoms does not even exist? Since a systematic approach was used to interpret the symptom scores in these DBPCFCs, it was possible to estimate that the risk of a false positive was lower than 20%, making it unlikely that all positives were false in that study. In this way it can be proven that adults can suffer from subjective symptoms caused by food.

Evaluations of new tests may also lack in validity. For instance, if a new test is investigated and it is shown to have a rather low agreement with the DBPCFC, it could still be a superior test if the disagreement between the 2 tests actually occurred in the patients where DBPCFC was incorrect. This problem may seem to be impossible to solve in the absence of a test to confirm the DBPCFC results, making the only possibility follow-up of patients to assess response to treatment and determine whether the effect of diet corresponds to the diagnosis. However, DBPCFC might actually be one of the few reference tests for which it is possible to make some accuracy calculations without knowing the true status of the patients. DBPCFC could be regarded as a cross-over study (or an n-1 trial) in a single patient, and this may allow some basic accuracy calculations to be made, as discussed in the next section. We strongly recommend performing such accuracy studies for different variables included in the DBPCFC. As discussed earlier, due to the risk and the fact that the DBPCFC is time consuming, several different attempts have been made to find new tests [30-37]. The question is whether these evaluations and conclusions are correct, since they are based on an imperfect reference test with an unknown accuracy. Much of the research into methods for the diagnosis of food allergy depends on the assumption that DBPCFC has a high accuracy. There have also been attempts to find adequate thresholds of existing tests that would make DBPCFC superfluous. Among these, decision points for predicting a positive oral food challenge have been established for specific serum IgE and SPT [29-33]. However, it has been shown that only a very limited number of patients exceed these thresholds for making DBPCFC superfluous. This is also valid, for combinations of tests, such as SPT and atopy patch test. Nevertheless, these results assume that DBPCFC is highly accurate, and under that assumption DBPCFC is for the moment the best available test.

Suggested Accuracy Calculations

The sensitivity of a diagnostic test is defined as the proportion of positive tests among individuals who actually have the target disorder. Similarly, the specificity is the proportion of negative tests among the individuals without the target disorder. There is, however, no consensus about how the results of DBPCFC should be interpreted in terms of when it should be considered positive or negative. In fact, there are a number of different possibilities, as discussed below. In the calculations given below, we assume that the results between different provocations are independent.

Dichotomization Strategy

The DBPCFC includes a comparison between 1 active provocation and 1 placebo provocation. Each provocation is considered as positive or negative and the overall DBPCFC is considered positive only if the active provocation was positive and the placebo provocation was negative. Sensitivity of a provocation is defined as the probability of receiving a positive active provocation given that the patient is truly food allergic. Specificity of a provocation is defined as the probability of receiving a negative result for a provocation (no matter if it is active or placebo), given that the patient is truly not food allergic. Given these definitions the probability of receiving a false-positive DBPCFC (ie, false positive on active provocation and negative on placebo provocation) is (1-specificity) × specificity. In other words the specificity of a DBPCFC is 1-([1-specificity of a provocation] × specificity of a provocation). For instance, if the specificity of a provocation is 95% then the overall specificity of the DBPCFC is 1-([1-0.95] × 0.95) = 95.25%.

The overall sensitivity of DBPCFC is calculated as the probability of positive active provocation × probability of negative placebo provocation = sensitivity of a provocation × specificity of a provocation. The problem is that the sensitivity of a provocation is unknown (since we never know if a patient is truly food allergic or not). However if we assume that the sensitivity is, for instance, 90%, the overall sensitivity of the DBPCFC is 0.9 × 0.95 = 85.5%

In summary, with the dichotomization strategy the accuracy of the DBPCFC is as follows:

Specificity = 1-([1-specificity of a provocation] ×specificity of a provocation)

Sensitivity = sensitivity of a provocation × specificity of a provocation

The specificity can be estimated by finding the specificity of a provocation through empirical observation of the proportion of negative placebo provocations. The sensitivity cannot be calculated without making assumptions. Naturally it is a good idea to first reach a consensus regarding when a provocation should be considered as positive or negative. For instance, whether a provocation including spirometry should be considered as positive if the values decrease by 10%, 15%, or 20%. The choice of threshold affects the specificity of the provocation. To define a positive reaction may be even more difficult when there is a lack of precise measurements, such as in the case of judging flush.

Difference Strategy

Instead of dichotomizing each provocation as either positive or negative, it is possible to directly use the difference between the observations received on active and placebo provocation. For instance, imagine that we use a spirometry test and forced expiratory volume in 1 second as a variable and that a reduction of 15% is considered as a positive provocation. If the reduction after active provocation is 16% and after placebo 7%, the DBPCFC should be considered as positive if the dichotomization strategy is used. However, the difference strategy means that we should consider the difference (16% – 7% = 9%) instead. This is actually more statistically correct, since it is a more efficient use of figures. Note that the 2 approaches may reach different conclusions; if the reduction on active provocation is 16% but reaction on placebo is 14% the dichotomization strategy yields a positive DBPCFC while the difference approach implies a negative DBPCFC.

If the difference strategy is used, the specificity of a DBPCFC is defined as the probability of receiving a difference above the chosen threshold by chance. In other words, if we estimate the variability of reactions seen on placebo (eg, estimate the SD), we could also find the SD of differences between 2 placebo provocations calculated as SD $\times\sqrt{2}$. Furthermore, if the differences seem to be fairly Gaussian and if we choose a threshold equal to 1.64 × SD × $\sqrt{2}$ we will receive a specificity of around 95%. Other specificities are found by changing the value 1.64.

This is just according to standard statistical theory for describing the distribution of a difference between 2 observations. To calculate the sensitivity of the DBPCFC we must assume the expected reaction; ie, the expected difference between active provocation and placebo, given that the patient is truly food allergic. This is also standard statistical theory analogous to a power calculation.

Nonparametric Strategy

If the reactions to food allergy are subjective the use of repeated provocations is recommended. In the Gothenburg center we apply 5 provocations when patients only present subjective symptoms. We randomly choose a sequence of provocations including 2 or 3 active and the corresponding 2 or 3 placebos. When the symptoms are subjective (ie, the variables used to characterize the reactions are qualitative data), it is reasonable to use nonparametric statistical analyses. In the Gothenburg centre, a DBPCFC is defined as positive if the mildest reactions seen on an active provocation are of a higher magnitude than the reactions seen on the worst placebo provocation. In a DBPCFC using 5 provocations this approach is analogous to a Mann–Whitney *U* test, that is, to rank all provocations and to see if all placebos are milder than all actives and, if that is the case, to regard the DBPCFC as positive. It is possible to show that this approach gives a specificity for the DBPCFC equal to 95%. The sensitivity of DBPCFC in this situation depends on the probability that an active provocation gives worse reactions than a placebo. For a given sensitivity and following the dichotomization strategy described above, it is possible to calculate the corresponding probability that an active provocation gives worse reactions than a placebo, as required to calculate the power of a Mann–Whitney *U* test. When we did this, we found that the nonparametric approach including 5 provocations had the same sensitivity as using 2 provocations with the dichotomization approach for objective symptoms. Thus, it is possible to study subjective symptoms with the same accuracy as objective symptoms, just by using repeated provocations.

The options discussed above are not exhaustively described in terms of statistical details. However, we hope that this could be a step towards formalization of how to interpret results, how to define positive and negative provocations, and DBPCFC in general. Furthermore, we hope that accuracy studies including statistical calculations of specificity and perhaps also scenarios regarding sensitivity will form part of future developments in the diagnosis of food allergy.

Other Important Methodological Issues

As a rigorous diagnostic study may be time consuming and expensive, it is quite common to evaluate several different new potential tests in the same study. This implies a classical statistical problem of multiplicity in which significant results can be obtained by chance when several analyses are performed. For instance, if a study includes 20 different new potential tests, 1 is expected to show significant discriminatory ability just by chance. Furthermore, it is common for the threshold level not to be stated in advance, and this gives the investigator the opportunity to choose the most appropriate level adjusted for that specific observed data set. In such an approach, the diagnostic precision may be over optimistic [48-49]. In the case of clinical trials, the ICHE-9 guidelines recommend that studies be divided into exploratory and confirmatory [50]. We suggest the same categorization for diagnostic studies. In an exploratory study, several new tests could be evaluated and elaborations with threshold levels could be performed to optimize the diagnostic accuracy. However, such a study must be followed by a confirmatory study where the number of test variables is limited (preferably only 1) and where threshold levels are defined in advance.

In most situations, the diagnostic process is multivariate, and therefore, it is also reasonable to carry out multivariate diagnostic research; that is, to evaluate whether the test contributes additional information not obtained by physical examination, anamnesis, and other tests [14]. It is suggested that such research questions should be described as diagnostic research, while univariate evaluations of a single test should be described as test research [51-52].

In food allergy, the diagnostic work-up includes medical history, several different tests, and open challenges before the DBPCFC. When evaluating the DBPCFC or another new test, it is advisable to consider a multivariate assessment including adequate analysis, such as multivariate logistic regression.

Conclusions

We have presented an overview of several important methodological issues that we recommend be considered in forthcoming studies in order to ensure the highest validity and clinical usefulness possible. One of the most crucial issues in diagnostic accuracy studies is the choice of reference test. DBPCFC has been used as the reference in studies addressing the diagnosis of food allergy. Consequently, much of the scientific development in this area relies on a belief that the DBPCFC provides a definitive diagnosis. For instance, while a number of attempts have been made to find tests that are simpler to use in clinical practice, the tests have been shown to lack accuracy compared to the DBPCFC, assuming that the result of DBPCFC is definitive. However, this might not be true, since the tests being evaluated may have disagreed with DBPCFC in the cases where DBPCFC was in fact wrong. Naturally, this is a hypothetical discussion, but it clearly demonstrates the need for further evaluation of the accuracy of DBPCFC.

We also recommend that strong efforts should be made to improve standardization of DBPCFC in clinical practice, and especially how to interpret the outcome. It is a little bit surprising that existing guidelines do not include discussions of thresholds and how to define positive and negative provocations in DBPCFC. Finally, studies are warranted to evaluate the accuracy of controlled oral food challenges.

References

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA. 1995;274:645-51.
2. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of Variation and Bias in Studies of Diagnostic Accuracy – a systematic review. Ann Intern Med. 2004;140:189-202.
3. Knottnerus JA, Van Weel C, Muris JWM. Evaluation of diagnostic procedures. BMJ. 2002;324:477-80.
4. Sacket DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539-41.
5. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ. 2002;324:669-71.
6. Knottnerus JA editor. The evidence base of clinical diagnosis. London: BMJ books; 2002.
7. Pepe M. Statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2004.
8. Jaeschke R, Guyatt GH, Sacket DL. Users' guides to the medical literature III. How to use an article about a diagnostic test. Are the results of the study valid? JAMA. 1994;271:389-91.
9. Van Der Schouw YT, Verbeek ALM, Ruijs JHJ. Guidelines for the assessment of new diagnostic tests. Invest Radiol. 1995; 30:334-40.
10. Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health. 2002;56:337-8.
11. The STARD Statement for Reporting Studies of Diagnostic Accuracy. Available from: http://www.consort-statement.org/stardstatement.htm
12. Solberg HE. Approved recommendation (1986) on the Theory of Reference Values. Part 1. The Concept of Reference Values. J Clin Chem Clin Biochem. 1987;25:337-42.
13. Petitclerc C, Solberg HE. Approved recommendation (1987) on the Theory of Reference Values. Part 2. Selection of Individuals for the production of reference values. J Clin Chem Clin Biochem. 1987;25:639-44.
14. Solberg HE, Petitclerc C. Approved recommendation (1988) on the Theory of Reference Values. Part 3. Preparation of Individuals and Collection of Specimens for the production of Reference Values. J Clin Chem Clin Biochem. 1988;26:593-98.

15. Solberg HE, Stamm D. Approved recommendation on the Theory of Reference Values. Part 4. Control of Analytical Variation in the Production, Transfer and Application of Reference Values. Eur. J Clin Chem Clin Biochem. 1991;29:531-35.
16. Solberg HE. Approved recommendation (1987) on the Theory of Reference Values. Part 5. Statistical Treatment of Collected Reference Values. Determination of Reference Limits. J Clin Chem Clin Biochem. 1987;25:645-56.
17. Dybkaer R, Solberg HE. Approved recommendation (1987) on the Theory of Reference Values. Part 6. Presentation of Observed Values Related to Reference Values. J Clin Chem Clin Biochem. 1987;25:657-62.
18. The declaration of Helsinki. Available from: http://www.wma.net/e/ethicsunit/helsinki.htm
19. Bock SA. Food challenges in the diagnosis of food hypersensitivity. In: De Weck AL, Sampson HA, editors. Intestinal immunology and food allergy. New York: Raven Press; 1995. p.105-17.
20. Bindslev-Jensen C. Standardization of double-blind, placebo-controlled, food challenges. Allergy. 2001;56 Suppl 67:S83-5.
21. Briggs D, Aspinall L, Dickens A, Bindslev-Jensen C. Statistical model for assessing the proportion of subjects with subjective sensitisations in adverse reactions to foods. Allergy. 2001;56 Suppl 67 :S83-5.
22. Bindslev-Jensen C, Briggs D, Osterballe M. Can we determine a threshold level for allergenic foods by statistical analysis of published data in the literature? Allergy. 2002;57:741-6.
23. Niggemann B, Wahn U, Sampson HA. Proposals for standardization of oral food challenge tests in infants and children. Pediatr Allergy Immunol. 1994;5:11-13.
24. Bindslev-Jensen C, Ballmer-Weber BK, Bengtsson U, Blanco C, Ebner C, Hourihane J, Knulst A C, Moneret-Vautrin D A, Nekam K, Niggemann B, Osterballe M, Ortolani C, Ring J, Schnopp C, Werfel T. Standardization of food challenges in patients with immediate reactions to foods –position paper from the European Academy of Allergology and Clinical immunology. Allergy. 2004;59:690-97.
25. Niggemann B, Beyer K. Diagnostic pitfalls in food allergy in children. Allergy. 2005; 60:104-107.
26. Vila L, Sanz ML, Sánchez-Lópes G, García-Avilés C, Diéguez I. Variation of serum eosinophil cationic protein and tryptase, measured in serum and saliva, during the course of immediate allergic reactions to foods. Allergy. 2001;56:568-72.
27. Magnusson J, Gellerstedt M, Ahlstedt S, Andersson B, Bengtsson, Telemo E, Hansson T, Peterson C.G. A kinetic study in adults with food hypersensitivity assessed as eosinophil activation in fecal samples. Clin Exp Allergy. 2003;33:1052-1059.
28. Roehr C, Reibel S, Ziegert M, Sommerfeld C, Wahn U, Niggemann B. Atopy patch test together with level of specific IgE reduces the need for oral food challenges in children with atopic dermatitis. J Allergy Clin Immunol. 2001;107:548-553.
29. Sampson H.A. Utility of food-specific IgE concentrations in predicting symptomatic food allergy. J Allergy Clin Immunol. 2001;107:891-896.
30. Sampson H.A. Food allergy – accurately identifying clinical reactivity. Allergy. 2005:60 Suppl 79:S19-24.
31. Celik-Bilgili S, Mehl A, Verstege A, Staden U, Nocon M, Beyer K, Niggemann B. The predictive value of specific immunoglobulin E levels in serum for the outcome of oral food challenges. Clin Exp Allergy. 2005;35:268-73.
32. Verstege A, Mehl A, Rolinck-Werninghaus C, Staden U, Nocon M, Beyer K, Niggemann B. The predictive value of the skin prick test wheal size for the outcome of oral food challenges. Clin Exp Allergy. 2005;35:1220-6.
33. Niggemann B, Rolinck-Werninghaus C, Mehl, A, Binder C, Ziegert M, Beyer K. Controlled oral food challenges in children - when indicated, when superfluous? Allergy. 2005;60:865-70.
34. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005; 51(8):1335-41.
35. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-30.
36. Begg CB. Biases in the assessment of diagnostic tests. Statistics Med. 1987;6:411-23.
37. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39:207-15.
38. Knottnerus JA, Leffers JP. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epid. 1992;45:1143-54.
39. Diamond GA. Selection bias and the evaluation of diagnostic tests: a metadissent. J Chron Dis. 1986;39:359-60.
40. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med. 1992;117:135-140.
41. Mulherin SA, Miller CW. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137:598-602.
42. Moons KG, Van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8:12-7.
43. Fraser CG, Harris EK. Generation and application of data in biological variation in clinical chemistry. Crit Rev Clin Lab Sci. 1989;27:409-37.
44. Vohlonen I, Terho EO, Koivikko A, vanto T, Holmen A, Heinonen OP. Reproducibility of the skin prick test. Allergy. 1989;44:525-31.
45. Gellerstedt M, Magnusson J, Gråsjö U, Ahlstedt S, Bengtsson U. Interpretations of subjective symptoms in double blind placebo controlled food challenges – Inter-observer-relibility. Allergy. 2004;59:354-6.
46. Bindslev-Jensen C. Food allergy: A diagnostic challenge. Curr Probl Dermatol. 1999;28:74-80.
47. Hawkins D, Garrett M and Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat. Med. 20:1987-2001.
48. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515-24.
49. Harrell FE JR, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
50. International Conference On Harmonisation Steering Committee. ICH harmonised tripartite guideline. General considerations for clinical trials. Available from: http://www.ich.org/MediaServer:jser?@_ID=484&@_MODE=GLB

51. Moons KG, Van Es GA, Michel BC, Buller HR, Habbema JD, Grobbee DE. Redundancy of single diagnostic test evaluation. Epidemiology. 1999;10:276-81.

52. Moons KG , Biesheuvel CJ, Grobbee DE. Test research vs diagnostic research. Clinical Chemistry. 2004;50:473-76

Manuscript received January 14, 2007; accepted for publication April 30, 2007.

Martin Gellerstedt

University West
461 86 Trollhättan
Sweden
E-mail: martin.gellerstedt@hv.se