Accepted: 30 September 2017

DOI: 10.1111/pai.12811

ORIGINAL ARTICLE

Food Allergy

Physician's appraisal vs documented signs and symptoms in the interpretation of food challenge tests: The EuroPrevall birth cohort

Linus B. Grabenhenrich[1,2] | Andreas Reich[3] | Doreen McBride[4] | Aline Sprikkelman[5] | Graham Roberts[6] | Kate E. C. Grimshaw[7] | Alessandro G. Fiocchi[8] | Photini Saxoni-Papageorgiou[9] | Nikolaos G. Papadopoulos[10] | Ana Fiandor[11] | Santiago Quirce[11] | Marek L. Kowalski[12] | Sigurveig T. Sigurdardottir[13] | Ruta Dubakiene[14] | Jonathan O. B. Hourihane[15] | Leonard Rosenfeld[16] | Bodo Niggemann[16] | Thomas Keil[17] | Kirsten Beyer[16]

[1]Department for Infectious Disease Epidemiology, Robert Koch Institute, Berlin, Germany

[2]Department of Dermatology, Venerology and Allergology, Charité - Universitätsmedizin Berlin, Berlin, Germany

[3]German Rheumatism Research Center, Berlin, Germany

[4]HTA Health Economics Strategy and Research, RTI International, Manchester, UK

[5]Department of Pediatric Pulmonology & Pediatric Allergology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[6]University Child Health, University of Southampton, Southampton, Hampshire, UK

[7]Clinical Experimental Sciences, University of Southampton, Southampton, Hampshire, UK

[8]Department of Paediatrics, Paediatric Hospital Bambino Gesù, Rome, Italy

[9]Department of Allergy and Clinical Immunology, 2nd Pediatric Clinic, National & Kapodistrian University of Athens, Athens, Greece

[10]Division of Infection Immunity & Respiratory Medicine, University of Manchester, Manchester, UK

[11]Alergia Infantil, Hospital Universitario La Paz, Madrid, Spain

[12]Department of Immunology, Rheumatology and Allergy, Medical University, Lodz, Poland

[13]Department of Immunology, Landspitali - The National University Hospital of Iceland, Reykjavik, Iceland

[14]Faculty of Medicine, Vilnius University, Vilnius, Lithuania

[15]Paediatrics and Child Health, University College, Cork, Ireland

[16]Department of Paediatric Pneumology and Immunology, Charité - Universitätsmedizin Berlin, Berlin, Germany

[17]Institute for Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin, Germany

Correspondence
PD Dr. Linus B. Grabenhenrich, MD, MPH, Department for Infectious Disease Epidemiology, Robert Koch Insitute, Berlin, Germany.
Email: GrabenhenrichL@rki.de

Abstract

Background: Blinded food challenges are considered the current gold standard for the diagnosis of food allergies. We used data from a pan-European multicenter project to assess differences between study centers, aiming to identify the impact of subjective aspects for the interpretation of oral food challenges.

Methods: Nine study centers of the EuroPrevall birth cohort study about food allergy recruited 12 049 newborns and followed them for up to 30 months in regular intervals. Intensive training was conducted and every center visited to ensure similar handling of the protocols. Suspected food allergy was clinically evaluated by double-blind, placebo-controlled food challenges using a nine dose escalation protocol. The primary challenge outcomes based on physician's appraisal were compared to documented signs and symptoms.

Results: Of 839 challenges conducted, study centers confirmed food allergy in 15.6% to 53.6% of locally conducted challenges. Centers reported 0 to 16 positive placebo challenges. Worsening of eczema was the most common sign when challenged with placebo. Agreement between documented objective signs and the challenge outcome assigned by the physician was heterogeneous, with Cohen's kappa spanning from 0.42 to 0.84.

Conclusions: These differences suggest that the comparison of food challenge outcomes between centers is difficult despite common protocols and training. We recommend detailed symptom assessment and documentation as well as objective sign-based challenge outcome algorithms to assure accuracy and comparability of blinded food challenges. Training and supervision of staff conducting food challenges is a mandatory component of reliable outcome data.

KEYWORDS
data collection, decision-making, diagnostic techniques and procedures, food hypersensitivity, observer variation

1 | INTRODUCTION

Food allergy (FA) appears in diverse clinical patterns, typically involving the cutaneous, gastrointestinal, respiratory and cardiovascular systems.[1] Besides observable clinical signs, many patients and parents also report subjective symptoms. Infants may present with being uncomfortable or crying and preschool children may present with food refusal, unable to adequately express specific symptoms. A causal link to a trigger food is usually suspected when signs or symptoms occur within 2 hours of ingestion but delayed appearance is sometimes observed, for example worsening of eczema and gastrointestinal symptoms. The heterogeneity of FA impedes the development of a simple, comprehensive diagnostic workup.[2-5]

Clinical evaluation of FA is usually set in motion based on a suggestive medical history, sometimes complemented through a prospective dietician-supervised elimination diet. When the diagnosis is based only on self-reported symptoms or objective signs, a high number of healthy individuals are labeled food allergic.[6,7] Objective assessment of sensitization (eg, skin prick test, allergen-specific immunoglobulin E) is considered to be the first step toward a more objective case definition,[1,8] but only challenge testing can verify the etiologic role of a suspected food.[9] Current guidelines describe double-blind, placebo-controlled food challenges (DBPCFC) as the highest diagnostic standard.[4,8,10]

Variability may be introduced at the level of an individual physician's appraisal of signs and symptoms during food challenges, especially as those reported by food-allergic patients are expected to overlap with those of healthy individuals to a certain degree. A *permissive* decision point should be chosen to miss only a small number of possible food allergies, including mild types, but this may result in falsely labeling healthy individuals as food allergic, leading to unnecessary restriction of nutrition and to faulty self-perception of FA status. This may be an appropriate trade-off in clinical settings to secure the diagnosis of potentially life-threatening FA, but in research, observational and interventional, choosing a *restrictive* decision point based on more objectively measurable signs or symptoms would reduce the number of false positives and would strengthen comparability of data between study physicians (Figure 1).

The impact of personal experience and subjective appraisal of the clinical appearance on the diagnostic interpretation of blinded food challenges has rarely been examined.[11-13] Using data from single-protocol DBPCFCs conducted within the multicenter EuroPrevall birth cohort,[14-16] we aimed to compare challenge outcomes as defined by the experienced and trained study physicians with those based on detailed documented signs and symptoms. Our goal was to identify areas, which could be improved further to support comparability, including techniques used for challenge documentation and interpretation, and diagnostic algorithms to improve the current gold standard for a robust diagnosis of FA.

2 | METHODS

2.1 | Setting and participants

The EuroPrevall birth cohort set out to estimate the frequency and factors influencing the onset and duration of FA in 9 study centers in 9

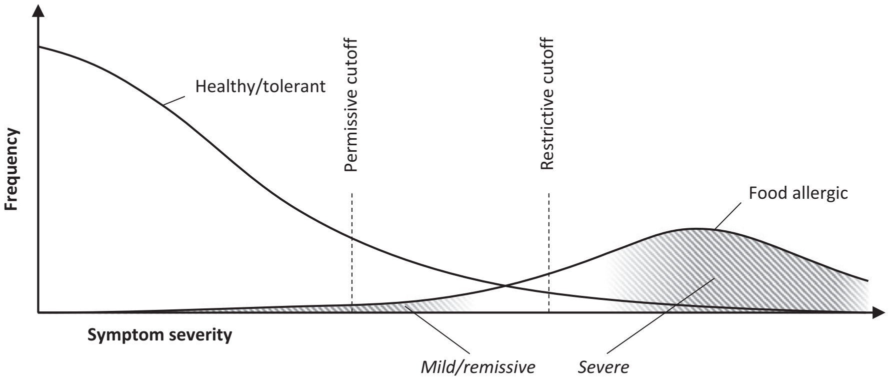

FIGURE 1 Hypothetical distribution of symptom severity upon double-blind, placebo-controlled food challenge in a preselected sample of preschool children evaluated for suspected food allergy matching eligibility criteria (eg, indicative history, specific sensitization), stratified by disease status (healthy vs food allergic)

different European countries. This initial phase of the cohort ran from birth to 30 months of age. Detailed methods have been published previously.[17] In short, 12 049 healthy newborns from the general population were enrolled before or shortly after birth and regularly followed in 6 month intervals, collecting data including dietary habits and other exposures. Parents were instructed to report to their study center immediately upon suspected FA or developing eczema. Additionally, interviews were conducted at 12, 24 and 30 months of age to screen for unrecognized signs or symptoms of food allergy. For each child invited to the center and two age-matched healthy controls per symptomatic child, we performed skin prick tests (SPTs) and measured specific immunoglobulin E (sIgE) antibodies in serum against six core allergens relevant in childhood (ie, cow's milk, hen's egg, wheat, soy, peanut, fish), plus suspected other food allergens. The decision to perform a DBPCFC was based on a positive test for allergic sensitization (ie, SPT wheal ≥3 mm or sIgE ≥0.35 kU/L) without currently eating the food, immediate objective clinical signs and symptoms, subjective reactions upon repeated exposures or clear improvement under elimination diet. Food challenges were performed in the participating clinics, supervised by trained physicians, and some centers asked families to stay overnight. Delayed symptoms were considered up to 48 hours after the challenge. Participants with confirmed FA were rechallenged after 12 months and, if still eligible, after 24 months of the initial diagnosis.

2.2 | Food challenges

The unit of observation for this analysis was a complete challenge including one food (verum) and a corresponding placebo control day. A single placebo day may have served as a control for more than one food in question. Two challenge days were randomly allocated to test food or placebo. Challenge and placebo days were at least 48 hours apart. Children were fed 9 incremental doses in 20 minute intervals under clinical supervision.[14] The procedure was stopped at the discretion of the supervising physicians. All physicians were trained in the food challenge protocol for this study to harmonize the identification of objective signs and symptoms warranting the discontinuation of the challenge. However, as food allergy has very diverse appearances, it was not possible to formally define all indications for stopping the challenge, in particular in light of the patient's safety. The assessment was unblinded after completion of the last challenge day.

In this analysis, we compared three different challenge-based definitions of food allergy, described in the following paragraphs.

2.3 | Physician's judgment of challenge outcome

For the first definition (physician's judgment), study physicians assigned outcomes (positive, negative) for each challenge day and then concluded an overall judgment after unblinding. This overall conclusion was the first definition of food allergy used. Patients were judged to be clinically tolerant (both days negative), allergic (test food positive, placebo negative), placebo reactors (test food negative, placebo positive), or inconclusive with regard to food allergy (both days positive).

2.4 | Restrictive cutoff for challenge outcome

For the other two definitions of food allergy, clinical observations were recorded through a standardized record form with separate sections for each challenge step recording immediate and delayed (≥2 hours) reactions. Besides skin assessment (SCORAD [18]) and vital parameters before and after the challenge, 19 specific signs and symptoms were collected as dichotomous traits (present or absent). Two different cutoff criteria were used to derive sign- and symptom-based challenge outcomes. After the judgment of the study physician was recorded, the *restrictive* cutoff (second definition of food allergy) to call a challenge positive was derived, limited to immediate urticaria, angioedema, flush, emesis, diarrhea, respiratory symptoms, immediate or late worsening of eczema with an objective SCORAD increase ≥10, and cardiovascular symptoms (never observed in this population).

2.5 | Permissive cutoff for challenge outcome

The *permissive* cutoff (third definition of food allergy) additionally included reactions occurring more than 2 hours after the challenge (called delayed) and less pronounced worsening of eczema (SCORAD increase of 5 or more).

2.6 | Statistical methods

Calculations were performed using SAS 9.4 (SAS Institute Inc., Cary, NC, USA). Missing data was rechecked against the initial study documentation, and only challenges with a known food item and study physician's final outcome decisions (first definition of food allergy) were used in this analysis. Agreement between sign- and symptom-based (second and third definitions) vs physician's appraisal (first definition, Cohen's Kappa coefficient [19]) was calculated only for subsamples large enough to report robust proportion estimates (20 + reactive challenges for a single center).

3 | RESULTS

3.1 | Challenge outcomes

A total of 839 valid food challenges (verum-placebo pairs) were conducted in the EuroPrevall birth cohort. Although study centers were similar in size (976 to 1570 participants), they reported widely differing numbers of procedures (7 to 219). Based on study physician's judgment (first definition of challenge-based food allergy), 317 (38.8%) challenges resulted in the diagnosis "allergic" due to a positive outcome on the verum day and a negative outcome on the placebo day. Cow's milk and hen's egg were the most frequent foods in question. Percentages of confirmed FA varied between centers (26.1% to 80.0% of conducted challenges). The proportion of allergic to challenged children was similar across different ages. Challenges with positive placebo day (placebo reactors and inconclusive food challenges) were unequally distributed between centers, with a maximum of 16 procedures in center C. Twenty-eight of all 45 (62.2%) challenges with a positive placebo day were related to cow's milk, with a trend toward younger ages (Table 1).

TABLE 1 Outcome of double-blind, placebo-controlled food challenges as stated by study physician, stratified by center, food item, and age

		Verum-placebo pairs	Reactive	Positive placebo day[a]
All challenges, n (%)		**839**	**317 (38.8)**	**45 (5.4)**
Country	A	139	56 (40.3)	15 (10.8)
	B	113	60 (53.1)	9 (8.0)
	C	219	78 (35.6)	16 (7.3)
	D	75	32 (42.7)	0 (0.0)
	E	120	50 (41.7)	3 (2.5)
	F	28	15 (53.6)	0 (0.0)
	G	96	15 (15.6)	1 (1.0)
	H	42	9 (21.4)	0 (0.0)
	I	7	2 (28.6)	1 (14.3)
Food	Cow's milk	368	109 (29.6)	28 (7.6)
	Hen's egg	288	133 (46.2)	6 (2.1)
	Wheat, Soy, Fish, Peanut, Tree nuts	160	69 (43.1)	6 (3.8)
	Other allergens	23	6 (26.1)	5 (21.7)
Age	0 - 11 mo	246	100 (40.7)	25 (10.2)
	12 - 23 mo	369	131 (35.5)	12 (3.3)
	24 mo and older	224	86 (38.4)	8 (3.6)

[a] with or without a positive verum day.

3.2 | Challenge signs and symptoms

A total of 334 of 839 (39.8%) verum (test food) challenge days were stopped before starting the final dose, of which most instances were judged positive by physicians. Urticaria (30.9%), flush (29.4%), and respiratory signs or symptoms (36.8%) were the most frequent reasons to stop challenges at lower doses (after step 1 to 4), accompanied by subjective gastrointestinal symptoms in 35.3% (not always considered as stop criterion on its own, Table 2). Food challenges were commonly stopped later (after step 5 to 8) because of urticaria and flush (33.8% and 18.4%), usually with no indication of respiratory or subjective gastrointestinal symptoms. Emesis and nasal/ophthalmic signs and symptoms appeared with increasing dose steps. Worsening of eczema was commonly reported (12.0%) but was only considered a stop criterion when supported by an objective SCORAD increase≥10. After completing the final dose (step 9), early (<2 hours) objective skin signs and emesis were among the most commonly documented. Delayed reactions (≥2 hours) included diarrhea, subjective gastrointestinal symptoms, and often worsening of eczema (without documented SCORAD, as parents reported it from home). Of the 101 placebo provocations, which did not reach the final dose (both, rated positive or negative), no clear sign or symptom was documented to why the procedure was stopped. This was likely due to the large amount [14] of test food (both for verum and placebo days) relative to children's age, as reported by study personnel. In patients who completed all placebo doses later rated as a positive challenge, emesis, diarrhea, flush and worsening of eczema were reported after the final placebo dose commonly, both as immediate- and delayed-type reactions.

3.3 | Recalculated outcomes based on signs and symptoms

All challenge outcomes were later recalculated based on objective challenge signs and symptoms. Using criteria as already defined within the study protocol (here called *restrictive* cutoff, second definition of food allergy), the number of reactive challenges was lower than when based on physician's judgment (252 vs 317, 22% reduction). Comparison of centers revealed pronounced differences with a reduction of 53% (37 restrictive-diagnosed vs 78 physician-diagnosed in center C) compared to an increase of 40% (21 restrictive-diagnosed vs 15 physician-diagnosed) in center G. Including delayed objective signs and a lower SCORAD cutoff (increase ≥5) in the sign- and symptom-based outcome definition (here called *permissive* cutoff, third definition of food allergy), labeled 63 more challenges reactive, with a maximum in one center of 29 (center C, Figure 2). Looking at confirmed FA using a restrictive cutoff, 93 of

TABLE 2 Symptoms following highest administered dose, by challenge day (verum/placebo). All centers, all ages, all food items. Numbers represent single challenge days. Shading: symptoms accounted for as stop criteria used in symptom-based challenge outcome definition (light: permissive cutoff, dark: restrictive cutoff). The following symptoms were never reported and thus not shown here: Blisters in oral mucosa, dysphagia, blood pressure drop, and shock

	Verum day				Placebo day			
Dose-symptom interval	<2 h			≥2 h	<2 h			≥2 h
highest dose administered[b]	1.4	5.8	9	1.9	1.4	5.8	9	1.9
Number of challenges	(68)	(266)	(505)	(839)	(22)	(79)	(551)	(652)
Skin								
Urticaria	21	90	28	26	6	2	0	2
Angioedema	5	11	3	5	0	0	0	0
Flush	20	49	26	29	0	2	7	4
Eczema								
Any	7	32	27	81	2	1	11	18
Increased SCORAD ≥5	2	17	9	11	1	1	0	0
Increased SCORAD ≥10	1	10	5	2	1	0	0	0
Gastrointestinal								
Emesis	9	33	21	23	0	2	9	4
Diarrhea	1	7	6	41	0	1	1	6
Subjective (pain, nausea, OAS)	24	0	0	27	3	0	0	3
Respiratory, ENT								
Airways[a]	25	0	0	1	0	0	0	1
Nasal	3	15	5	5	0	0	0	0
Ophthalmic	1	14	5	6	0	0	0	0

OAS, oral allergy syndrome; ENT, Earnose and throat.

[a] Following a nine dose protocol as explained in(14).

[b] Bronchospasm, dyspnea, cough, laryngoedema.

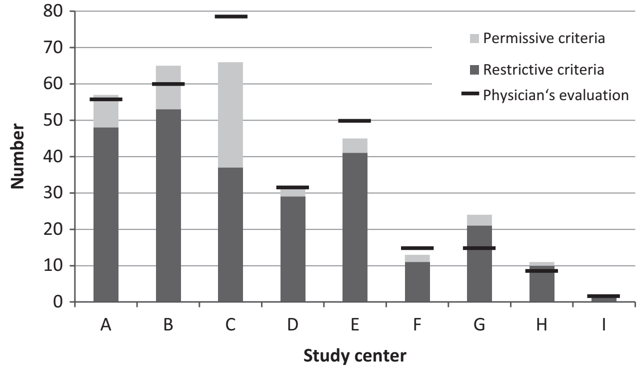

FIGURE 2 Number of reactive challenges per study center (of 9), based on symptom profile. Permissive cutoff including delayed reactions and worsening of eczema with a SCORAD increase ≥5, restrictive cutoff accounting only for early objective symptoms and worsening of eczema with SCORAD increase ≥10

252 (36.9%) challenges did complete all steps of the placebo day as would have been required by the study protocol. This occurred similarly at all ages. Centers varied with respect to not finishing placebo challenges (descending frequencies, center G 57.1, C 56.8, E 51.2%). The two cutoffs resulted in different numbers of reactive challenges across all ages (Figure S1).

3.4 | Sign- and symptom-based outcomes vs physician's appraisal

The agreement between sign- and symptom-based challenge day outcomes using the *restrictive* cutoff (considered as stopping criteria in the study protocol) and physician's judgment/diagnosis varied between study centers, with the lowest agreement in center C yielding a Cohen's kappa coefficient of 0.42, and highest agreement in center D (kappa 0.84). Higher degrees of agreement were achieved using *permissive* cutoff criteria, which were more similar between centers (range center C 0.74 to D 0.92, Figure 3).

4 | DISCUSSION

4.1 | Main results

There were differences between centers comparing physician's appraisal and sign- and symptom-based outcomes recorded during blinded food challenges of infants and young children up to the age of 2 years within the multicenter EuroPrevall birth cohort study. The agreement between the permissive cutoff and physician's appraisal was higher compared to the restrictive criteria, indicating a tendency for study physicians to apply a rather permissive decision threshold.

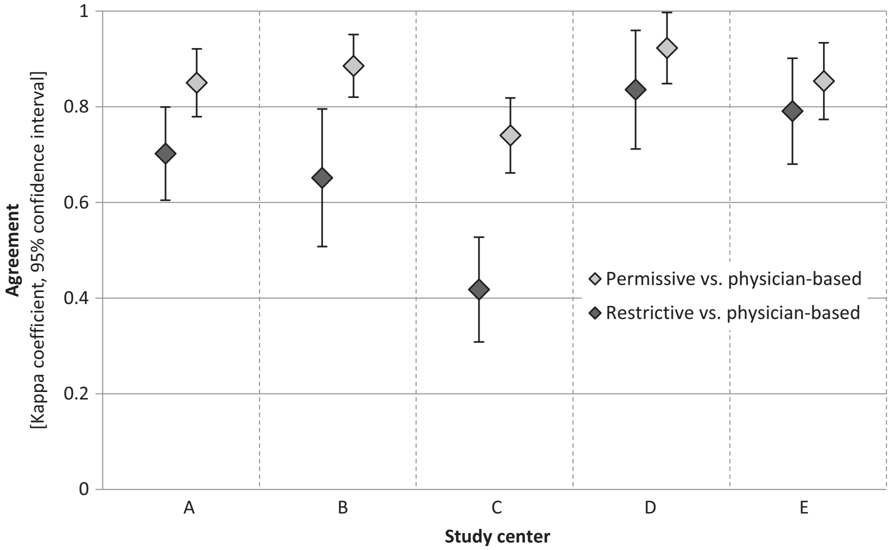

FIGURE 3 Agreement between study physician's judgment and symptom profile, using different symptom cutoffs (restrictive, permissive). Comparison of single day outcomes (test food and placebo). Asymptotic 95% confidence intervals, 4 of 9 study centers omitted due to low numbers of challenges

The wide range of positive challenge outcomes between centers (15.6% to 53.6%) might either be due to a real difference in disease incidence, unequal inclusion thresholds to perform a challenge testing, or may, at least in part, result from differences in documenting and interpreting signs and symptoms of oral food challenges. This emphasizes the need for standardization of all aspects of DBPCFCs including inclusion criteria, documentation of denial to participate, challenge conduct, and interpretation of the challenge outcome.

While the necessity of blinding in food challenges has been questioned (eg, [20]), the considerable numbers of signs and symptoms during placebo challenges seen in this study, especially delayed self-reported eczema (18 times on placebo vs 81 on test food days), demonstrate that blinding is imperative for accurate interpretation of food challenges. Interestingly, subjective gastrointestinal symptoms occurred almost only during the first lower doses, whereas subjective ENT symptoms and worsening of eczema were more common with higher doses of the tested allergen. This could be due to different mechanisms of action, be it psychologic or biologic. The frequent failure to reach the final dose during placebo challenges might be explained by the relatively large amounts of challenge agent used. Interestingly, a high number of placebo reactors were seen in the first year of life, as has been shown previously.[21] This finding stresses the need to perform blinded food challenges even in very young patients.

Detailed assessment and documentation of challenge signs and symptoms is a cornerstone of comparability, as seen by the difference between any eczema and SCORAD-scored eczema at the highest dose administered (on 66 vs 16 test food days). It is likely that grading of other signs and symptoms could further improve the accuracy of food challenges. Additionally, only three centers (A, C, F) recorded considerable numbers of subjective symptoms, supporting the need for a detailed assessment and documentation of these observations to be part of the challenge protocol. These details could include grading, measurement, or weighting to improve comparability of challenge results.

That judgment of symptoms and clinical signs always relies on individual experience and appraisal threatens the validity of comparisons between study centers and observers, indicated by the considerable differences of positive placebo challenges (0 to 16 per center) and variable agreement comparing physician-based vs objective sign-based challenge outcomes (Cohen's kappa spanning from 0.42 to 0.84, restrictive cutoff). In general, using a permissive cutoff yielded higher agreement with physician-based outcomes in all centers, highlighting the need for a unified, robust, and objective sign-based case definition for research.

4.2 | Recommendations

Development and standardization of current guidelines and challenge protocols [4,8] for the diagnosis of FA in the clinical setting are ongoing and should be promoted.[1,2,5] Their focus lies mainly on methodological aspects in light of their first priority, to rule out or confirm FA in real-life medical care settings. Consequently, looking at different steps from suspicion to confirmation of FA, blinding of challenges, detailed sign and symptom assessment, and standardized interpretation of challenge outcomes are usually neglected,[3,22] relying mainly on personal experience and individual judgment (Figure 4). When food challenges are used in research settings, these procedural aspects are likely to influence estimates of disease frequency and severity considerably and must not be ignored in study protocols. Here, comparability and restrictive case definitions outweigh the usual "don't-miss-any" approach, which is appropriate for individual care, where a false positive is a safer misclassification than a false negative.

As was done in this study, preparation and distribution of test food and placebo substrate should be centralized and off-site in research settings, ensuring a high degree of blinding. Unblinding must be delayed until after the challenge documentation has been closed and, as is suggested to assess allocation concealment in clinical trials, blinding success should be documented by assessing participant's and study personnel's guessed allocation of each of the challenge days.

Secondly, already proposed but usually not implemented,[5] all signs and symptoms should be quantified using appropriate measures such as size, distribution, and severity for skin manifestations beyond eczema, or amount and kind of vomit and diarrhea. Moreover,

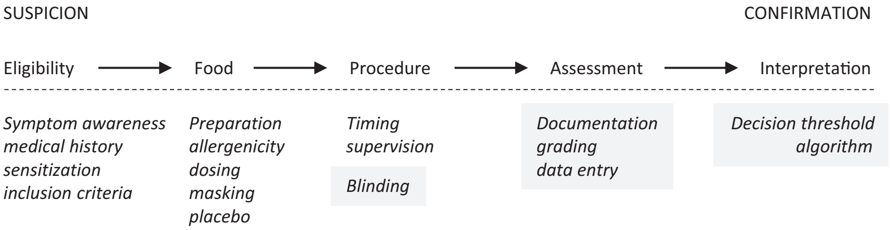

FIGURE 4 Blinded food challenge methodology. Highlighted aspects are commonly neglected in procedural guidelines/recommendations

there is an urgent need for a standardized assessment of potentially relevant gastrointestinal symptoms like colic, and general symptoms like crying and being uncomfortable. This is particularly needed for signs and symptoms commonly reported during placebo challenges and as delayed reactions (flush, urticaria, GI symptoms, [23]). Variation of clinical signs and symptoms, for example, worsening of eczema during or after the challenge day, can ideally be assessed by two independent physicians, the second blinded to judgment of the first, and not by parents alone. Peer-to-peer teaching and training of reaction assessment may shed light on under- or over-recognized signs and symptoms and improve comparability. As aimed for in this study, data entry for each challenge day should be closed before starting the next day. Additionally, documentation of challenge details would support any independent and objective consideration of challenge outcomes. These might include intentional and feasible protocol violations (eg, omission of the final dose), information about medical personnel (eg, level of experience), and the post-challenge period (eg, re-introduction of food, exact timing, and assessment of delayed reactions through professionals).

Thirdly, after closing data entry, a centralized evaluation scheme could assign the final challenge outcome based on recorded signs and symptoms, with the need to register its technical implementation as a medical device. Personnel on site should be asked to label observations they suspect to be causally linked to the ingested food, be it the allergen or placebo. Challenge outcome and day allocation (unblinding) could then be finally returned to the clinical site.

Finally, using a generic online platform for research as well as individual care settings may facilitate data entry, for example, ensuring that data entry for each challenge day was closed before starting the next day, and allowing on-time queries and electronic evaluation of challenge outcomes. Such an algorithm could be asked to return a binary decision (tolerant/reactive) using a rather loose cutoff with the intent to not miss any FA. It may at the same time report quantified severity of the reaction using other cutoffs, ultimately improving comparability between physicians, clinical sites, and countries.

Improving challenge guidelines is recommended to incorporate what we have demonstrated in this single-protocol, multicenter project, which could also be expected to be beneficial for regular patient care and other research settings.

4.3 | Strengths and limitations

As the current gold standard, blinded food challenges cannot be calibrated against another diagnostic test. Therefore, we used the ideal setting of a large single-protocol, multicenter research project to indirectly identify potential shortcomings of its diagnostic capabilities. Given stability of study personnel over time, heterogeneity of study centers in terms of initial experience with food challenges and different societal backgrounds between centers has allowed us to assess the influence of subjective (often undefined or not accessible) parameters, but we had no estimates for the individual experience of participating physicians. With the lack of comparable prior knowledge about disease frequency and distribution of potential subtypes of FA in participating countries, we were not able to directly separate true from factitious intercenter differences. We cannot prove that these differences indicate the influence of subjective parameters within this project alone or are rather due to possible disease heterogeneity, but with the procedural aspects identified here accounted for in future research, we will be closer to a true gold standard. Of note, the study was neither designed nor powered for the presented analyses.

4.4 | Conclusion

There is no methodology to assess the accuracy and other diagnostic characteristics of blinded food challenges directly. We demonstrated differences between centers of the multicenter EuroPrevall project in terms of overall reactivity to challenges, placebo reactions, and most importantly decision thresholds for assigning challenge outcomes based on physician's judgment. Despite using the same robust, highest standard challenge protocol, these discrepancies suggest there can still be a residual influence of subjective and other non-standardized parameters, threatening valid comparison of results between centers, if challenge outcome is not based on objective signs.

We recommend centralized provision of allergens for food challenges, implementation of detailed sign and symptom quantification, and timely documentation in standardized challenge record forms and that only pre-agreed sign- and symptom-based challenge outcomes derived by unified algorithms should be relied upon. These allow for continuous severity grading in addition to the usual dichotomous challenge outcome and provide valuable information for inter- and within-study comparisons. The school-age follow-up (iFAAM) of the EuroPrevall project implemented these recommendations using case report forms that are publicly available (CRFs, [24]). Accounting for these recommendations will further improve the diagnostic gold standard of blinded food challenges for food allergies.

ORCID

Linus B. Grabenhenrich http://orcid.org/0000-0002-9300-6625

Andreas Reich http://orcid.org/0000-0002-3729-0772

REFERENCES

1. Burks AW, Tang M, Sicherer S, et al. ICON: food allergy. *J Allergy Clin Immunol*. 2012;129:906-920.
2. Sampson HA, Gerth van Wijk R, Bindslev-Jensen C, et al. Standardizing double-blind, placebo-controlled oral food challenges: American Academy of allergy, asthma & immunology-European academy of allergy and clinical immunology PRACTALL consensus report. *J Allergy Clin Immunol*. 2012;130:1260-1274.
3. Nowak-Wegrzyn A, Assa'ad AH, Bahna SL, Bock SA, Sicherer SH, Teuber SS. Work group report: oral food challenge testing. *J Allergy Clin Immunol*. 2009;123:S365-S383.
4. Bindslev-Jensen C, Ballmer-Weber BK, Bengtsson U, et al. Standardization of food challenges in patients with immediate reactions to foods–position paper from the European academy of allergology and clinical immunology. *Allergy*. 2004;59:690-697.
5. Niggemann B, Beyer K. Diagnosis of food allergy in children: toward a standardization of food challenge. *J Pediatr Gastroenterol Nutr*. 2007;45:399-404.
6. Rona RJ, Keil T, Summers C, et al. The prevalence of food allergy: a meta-analysis. *J Allergy Clin Immunol*. 2007;120:638-646.
7. Zuidmeer L, Goldhahn K, Rona RJ, et al. The prevalence of plant food allergies: a systematic review. *J Allergy Clin Immunol*. 2008;121:e4.
8. Boyce JA, Assa'ad A, Burks AW, et al. Guidelines for the diagnosis and management of food allergy in the United States: report of the NIAID-sponsored expert panel. *J Allergy Clin Immunol*. 2010;126:S1-S58.
9. 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;60:865-870.
10. Niggemann B, Beyer K. Pitfalls in double-blind, placebo-controlled oral food challenges. *Allergy*. 2007;62:729-732.
11. van Erp FC, Knulst AC, Meijer Y, Gabriele C, van der Ent CK. Standardized food challenges are subject to variability in interpretation of clinical symptoms. *Clin Transl Allergy*. 2014;4:43.
12. Brand PL, Landzaat-Berghuizen MA. Differences between observers in interpreting double-blind placebo-controlled food challenges: a randomized trial. *Pediatr Allergy Immunol*. 2014;25:755-759.
13. Gellerstedt M, Magnusson J, Grajo U, Ahlstedt S, Bengtsson U. Interpretation of subjective symptoms in double-blind placebo-controlled food challenges - interobserver reliability. *Allergy*. 2004;59:354-356.
14. Keil T, McBride D, Grimshaw K, et al. The multinational birth cohort of EuroPrevall: background, aims and methods. *Allergy*. 2009;65:482-490.
15. Schoemaker AA, Sprikkelman AB, Grimshaw KE, et al. Incidence and natural history of challenge-proven cow's milk allergy in European children–EuroPrevall birth cohort. *Allergy*. 2015;70:963-972.
16. Xepapadaki P, Fiocchi A, Grabenhenrich L, et al. Incidence and natural history of hen's egg allergy in the first 2 years of life - the EuroPrevall birth cohort study. *Allergy*. 2015;71:350-357.
17. McBride D, Keil T, Grabenhenrich L, et al. The EuroPrevall birth cohort study on food allergy: baseline characteristics of 12,000 newborns and their families from nine European countries. *Pediatr Allergy Immunol*. 2012;23:230-239.
18. Sprikkelman AB, Tupker RA, Burgerhof H, et al. Severity scoring of atopic dermatitis: a comparison of three scoring systems. *Allergy*. 1997;52:944-949.
19. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
20. Venter C, Pereira B, Voigt K, et al. Comparison of open and double-blind placebo-controlled food challenges in diagnosis of food hypersensitivity amongst children. *J Hum Nutr Diet*. 2007;20:565-579.
21. Ahrens B, Niggemann B, Wahn U, Beyer K. Positive reactions to placebo in children undergoing double-blind, placebo-controlled food challenge. *Clin Exp Allergy*. 2014;44:572-578.
22. Jarvinen KM, Sicherer SH. Diagnostic oral food challenges: procedures and biomarkers. *J Immunol Methods*. 2012;2:30-38.
23. Niggemann B. When is an oral food challenge positive? *Allergy*. 2010;65:2-6.
24. Grabenhenrich LB, Reich A, Bellach J, et al. A new framework for the documentation and interpretation of oral food challenges in population-based and clinical research. *Allergy*. 2016;72:453-461.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Grabenhenrich LB, Reich A, McBride D, et al. Physician's appraisal vs documented signs and symptoms in the interpretation of food challenge tests: the EuroPrevall birth cohort. *Pediatr Allergy Immunol*. 2018;29:58–65. https://doi.org/10.1111/pai.12811