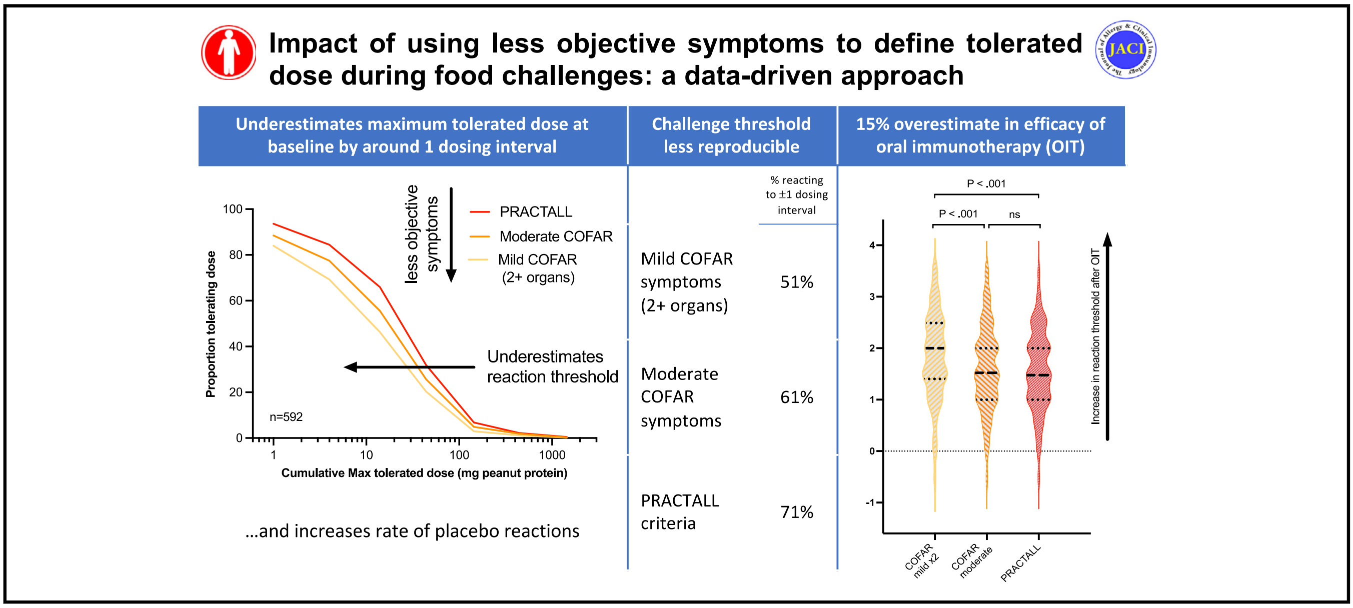
Impact of using less objective symptoms to define tolerated dose during food challenges: A data-driven approach

Paul J. Turner, FRCPCH, PhD,[a] Nandinee Patel, MRCPCH, PhD,[a] Katharina Blumchen, MD,[b] Stefanie Berkes, BSc,[a] Hugh A. Sampson, MD,[c] and Kirsten Beyer, MD[d] *London, United Kingdom; Frankfurt am Main and Berlin Germany; and New York, NY*

GRAPHICAL ABSTRACT



Impact of using less objective symptoms to define tolerated dose during food challenges: a data-driven approach

Underestimates maximum tolerated dose at baseline by around 1 dosing interval

Challenge threshold less reproducible

15% overestimate in efficacy of oral immunotherapy (OIT)

...and increases rate of placebo reactions

	% reacting to ±1 dosing interval
Mild COFAR symptoms (2+ organs)	51%
Moderate COFAR symptoms	61%
PRACTALL criteria	71%

Background: Food challenges (FCs) form the basis for assessing efficacy outcomes in interventional studies of food allergy; however, different studies have used a variety of similar but not identical criteria to define a challenge reaction, including subjective (nonobjective) symptoms occurring in a single-organ system as dose limiting.

Objective: Our aim was to undertake a secondary analysis of 4 interventional studies to assess the impact of using less objective criteria to determine challenge-stop on reaction thresholds and their reproducibility.

Methods: We analyzed individual participant data, including individual participant data meta-analysis, by using 3 different published challenge-stop criteria: (1) PRACTALL consesus criteria; (2) Consortium for Food Allergy Research version 3 (CoFAR v3) with at least 1 moderate- or severe-grade symptom; or (3) CoFAR v3 with at least 2 mild symptoms occurring in

From [a]the National Heart and Lung Institute, Imperial College London; [b]the Department of Paediatric and Adolescent Medicine, Pneumology, Allergology and Cystic Fibrosis, University Hospital Frankfurt, Goethe University; [c]the Division of Pediatric Allergy and Immunology, Jaffe Food Allergy Institute, Icahn School of Medicine at Mount Sinai, New York; and [d]the Department of Pediatric Respiratory Medicine, Immunology and Critical Care Medicine, Charité Universtãtsmedizin Berlin.

Supported by the UK Medical Research Council (award MR/S036954/1 [to P.J.T.]), National Institute for Health and Care Research/Imperial Biomedical Research Ccntre (to P.J.T.), and JM Charitable Foundation (to P.J.T.).

Disclosure of potential conflict of interest: P. J. Turner reports personal fees from the UK Food Standards Agency, Aimmune Therapeutics, Allergenis, Aquestive Therapeutics, and Novartis. K. Blumchen reports consulting for Aimmune Therapeutics, DBV Technologies, Bencard Allergie, Novartis; speakers bureau for Aimmune Therapeutics, DBV Technologies, HAL Allergy, Nutricia, ALK, Allergopharma, Nestle, and Novartis; and research grants from Aimmune Therapeutics, DBV Technologies, Novartis and Hipp, outside the submitted work. H.A. Sampson reports advisory board/consulting fees DBV Technologies, N-Fold, and Siolta Therapeutics; grant funding from the National Institutes of Health/National Institute of Allergy and Infectious Diseases; and royalties from Elsevier. K. Beyer reports advisory board/consulting fees from Aimmune Therapeutics, Allergy Therapeutics, Bencard, Danone, DBV, Hipp, Hycor, Jenapharm, Mylan/Meda Pharma/Mice, and Nestle; speakers bureau for Aimmune Therapeutics, ALK, Danone, Infectopharm, Med Update, Nestle, Nutricia; and research grants from Aimmune, ALK, Danone, DBV, Hipp, Hycor, Infectopharm, and Nutricia, outside the submitted work. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication August 21, 2022; revised November 29, 2022; accepted for publication December 20, 2022.

Corresponding author: Paul Turner, FRCPCH, PhD, National Heart & Lung Institute, Imperial College London, Norfolk Place, London, W2 1PG, UK. E-mail: p.turner@imperial.ac.uk.

0091-6749

https://doi.org/10.1016/j.jaci.2022.12.818

different organ systems. Reproducibility of challenge threshold was also assessed in participants undergoing subsequent repeat FCs.

Results: Four studies, with detailed challenge data from a total of 592 participants, were included. Applying CoFAR v3 definitions for dose-limiting symptoms resulted in an underestimate of reaction thresholds compared with those in PRACTALL (*P* < .001) that is equivalent to almost a single dosing increment when using a semi-log dosing regimen. Reproducibility was also reduced when applying CoFAR v3 (*P* < .001 [n = 223]). Using the least conservative interpretation of CoFAR v3 (≥2 mild symptoms occurring in different systems) resulted in a significant overestimate of 15% when assessing oral immunotherapy efficacy. Applying a data-driven minor modification to CoFAR v3 resulted in a new set of challenge-stop criteria with validity similar to that of PRACTALL but one that is simpler to implement and in which significant gastrointestinal discomfort with observable decreased activity remains a dose-limiting symptom.

Conclusion: The use of less objective symptoms to define challenge-stop compromises the reproducibility of the FC as a tool to assess efficacy outcomes in interventional studies, and potentially overestimates the efficacy of the intervention tested. (J Allergy Clin Immunol 2023;■■■:■■■-■■■.)

Key words: *CoFAR, eliciting dose, food challenge, peanut, PRACTALL, reproducibility, thresholds*

Abbreviations used

CoFAR: Consortium for Food Allergy Research
DBPCFC: Double-blind placebo-controlled food challenge
DLS: Dose-limiting symptoms
FC: Food challenge
OIT: Oral immunotherapy

Oral food challenges (FCs) are a key tool to evaluate the impact of interventions that might alter clinical reactivity in food allergy. Although oral immunotherapy (OIT) is becoming increasingly mainstream, there is still a need for further research to evaluate protocols that might reduce the adverse events associated with treatment. In addition, a number of biologics are currently under investigation in phase 3 clinical trials, either in isolation or in combination with allergen immunotherapy.

The gold standard for diagnosis of food allergy is the double-blind, placebo-controlled FC (DBPCFC). DBPCFCs are frequently used to determine the primary outcome in clinical trials; therefore, robust and standardized protocols are needed. The American Academy of Allergy, Asthma & Immunology–European Academy of Allergy and Clinical Immunology PRACTALL consensus report on standardizing DBPCFCs, published in 2012,[1] quickly gained international acceptance and has been widely used for over a decade to inform the execution of DBPCFCs including challenge-stop criteria. More recently, we evaluated the reproducibility of DBPCFCs to peanut and cow's milk, and found that intraindividual reaction thresholds can vary 1000-fold, although for the majority (70%-80%) of individuals, this variability is limited to a half-log change in threshold,[2,3] which is equivalent to a single dosing increment when using a semi-log dosing regimen as recommended by PRACTALL.

The PRACTALL consensus relies on a "traffic light system" to inform challenge-stopping signs or symptoms: typically, FCs are halted either with the occurrence of at least 1 significant objective symptom or with a combination of less objective or subjective symptoms or signs in more than 1 organ system. This is because relying on subjective symptoms increases the risk of a false-positive test result[4] and underestimates the apparent challenge reaction threshold.[1]

However, not all clinical trials have used PRACTALL consensus criteria to determine challenge-stop.[5] For this purpose, some studies have utilized less rigid criteria—often referred to as dose-limiting symptoms (DLS).[6,7] Conversely, some studies have used alternatives that better define (compared with PRACTALL) challenge-stop symptoms.[8,9] Chinthrajah et al recently proposed an updated framework (reproduced in Fig 1) as part of the updated Consortium for Food Allergy Research (CoFAR) Grading Scale for Systemic Allergic Reactions.[10] This includes the possibility of nonobjective, subjective symptoms in a single-organ system being considered dose limiting and leading to stopping an FC. For example, under the new proposal, isolated symptoms such as "throat tightness without hoarseness" or "more than a few areas of erythema" (but not widespread erythema) are sufficient to terminate a challenge.

In this analysis, we re-evaluated detailed challenge data available from clinical trials in participants with peanut allergy, applying both PRACTALL and the new CoFAR version 3 (CoFAR v3) criteria to assess the impact of using less objective symptoms to define challenge-stop criteria.

METHODS

Study selection

Having previously undertaken a systematic review of the literature to identify studies that used DBPCFC to evaluate clinical reaction thresholds to peanut, we included 3 studies in which detailed symptom data (relating to both subjective and objective symptoms at each FC dose) were available and individuals had undergone a further FC followings the initial DBPCFC (conducted according to the same protocol). Two were allergen immunotherapy trials,[11,12] while the third was a study in which adults with peanut allergy and a positive DBPCFC were randomized to undergo a repeat FC with or without cofactors (exercise, sleep deprivation).[13] For the purpose of this analysis, data were used from the baseline DBPCFC and nonintervention challenge (without a cofactor), which for the majority of participants was an open FC conducted according to an identical protocol. A fourth study evaluating OIT to peanut was also included.[14] This study did not use DBPCFC; instead, it utilized an open FC protocol with a dosing interval of 2 hours but was otherwise performed according to PRACTALL consensus criteria. Further details relating to the included studies can be found in Table I.

Data extraction and analyses

Analyses were planned prospectively. Data relating to the dose causing DLS were extracted in duplicate by using the following 3 definitions: (1) the dose causing challenge-stopping symptoms according to the PRACTALL consensus; (2) the lowest dose associated with the occurrence of at least 1 moderate- or severe-grade symptom defined according to the proposed CoFAR v3 criteria (see Fig 1), and (3) the lowest dose associated with the occurrence of at least 2 mild symptoms according to CoFAR v3 occurring in more than 1 organ system.

For all cohorts, the cumulative maximum tolerated dose was set as the dose given immediately before that causing DLS. Any discrepancies identified

	PRACTALL	COFAR v3	LEAP Study	Proposed aligned DLS
Suggested FC stopping criteria:	One red or 3 orange symptoms	One red or 2 orange symptoms	One red or 2 orange symptoms	One red or 2 orange symptoms from 2 distinct categories
SKIN				
Scratching	Mild/moderate incl. >2mins Continuous with excoriations	Occasional only Protracted	Scratching <3mins Scratching >3mins	Scratching <3mins Scratching >3mins
Erythema	Few faint areas <50% of body Generalised / >50%	Faint More than a few areas Pronounced / Generalised	Not widespread or itchy Widespread itchy rash	Few faint areas <50% of body Generalised / >50%
Hives/urticaria	<3 lesions ≥3 lesions	Few/localised Numerous	<3 lesions ≥3 lesions	<3 lesions ≥3 lesions
Angioedema	Mild lip Any significant	Mild lip Any significant	Any	Mild lip Significant lip or face edema
RESPIRATORY				
Rhinitis	Rare bursts, occasional sniffing Moderate, frequent Severe, persistent and/or long bursts	Any	<3mins >3mins	Mild, infrequent Moderate, frequent Severe, persistent
Eyes	Intermittent rubbing Continuous rubbing/reddening	Not specified	Rubbing <3mins Rubbing >3mins	Minimal reddening, occasional rubbing Rubbing >3mins Reddening without significant rubbing
Objective wheeze	Any	Any	Any	Any
PHARYNGEAL/LARYNGEAL				
Oral cavity	Itchy mouth	Not specified	Not specified	Oral itch
Throat	Itchy throat Persistent tightness or pain	Discomfort Tightness	Not specified	Itchy throat Persistent* tightness or painful swallow
Cough	>3 discrete episodes of throat clearing Frequent dry cough	Occasional Persistent	Repetitive	>3 discrete episodes of throat clearing/cough Frequent dry cough
Voice	Hoarseness/stridor	Hoarseness/stridor	Dysphonia/stridor	Hoarseness/stridor
GASTROINTESTINAL				
Abdominal discomfort	Mild nausea/pain ± ↓activity Moderate nausea/pain ± ↓activity Severe nausea/pain ± ↓activity	Mild nausea/pain ± ↓activity Persistent moderate nausea or pain with ↓activity	Mild nausea/pain Moderate nausea/pain Severe pain >3mins	Non-persisting nausea Mild pain ± ↓activity or persistent* nausea Persistent* moderate pain with ↓activity
Vomiting	1 episode 2+ episodes	Vomit due to gag Any other vomit	Any	Vomit due to gag Any other vomit
Diarrhea	1 episode 2+ episodes	Not specified	Any	1 episode 2+ episodes
CARDIOVASCULAR/NEUROLOGICAL				
	Weak/dizzy/tachycardia Hypotension Change in mental status / LOC	Clinically significant hypotension Change in mental status	Hypotension	Feeling weak, tired, upset/agitated Clinically significant hypotension Change in mental status

FIG 1. Comparison of challenge-stop criteria used in the literature for clinical trials of food allergy, and a proposed modification of COFAR v3 to better align DLSs with PRACTALL and those utilizing the framework used in the Learning Early about Peanut Allergy (LEAP) study. Any red symptom is generally considered to be dose limiting. Three orange symptoms are considered to imply a likely allergic reaction under PRACTALL, whereas 2 orange symptoms are considered to imply a likely allergic reaction under CoFAR v3 or the LEAP study DLS criteria. Green symptoms do not contribute toward the decision to terminate an FC. *Persisting symptom is defined as ongoing beyond the duration of a dosing interval (minimum 20 minutes).

between extracted data were resolved by discussion and/or by contacting authors or study sponsors for clarification.

Statistical analyses were conducted using GraphPad Prism, version 9.4.1 (GraphPad Software, San Diego, Calif), and R software, version 4.0.3. All statistical tests were 2 tailed, and a P value less than .05 was considered significant. Agreement was assessed by using a κ statistic with linear weighting.[15]

To assess the reproducibility of challenge thresholds within individuals over time, we extracted individual participant data on all individuals who underwent a further FC following their initial challenge (conducted according to the same protocol) without undergoing any intervention. The log fold change in reaction threshold for each subject was calculated, and the distribution of these data was assessed using violin plots. A similar process was followed to evaluate log fold change in reaction threshold in those participants who underwent OIT in 2 of the 4 included trials.[12,14]

Finally, we evaluated the impact of various modifications to the CoFAR v3 framework for defining challenge-stop in terms of reaction thresholds, reproducibility, and log fold change in reaction threshold in OIT-treated participants compared with the PRACTALL criteria. Normality of distribution was assessed by using the D'Agostino-Pearson test, and the data were then used for individual participant data meta-analysis. Rates were pooled across studies by using a generalized linear mixed model in R (metaprop function, metafor package, and logit transformation with a random effects model for the summary estimate, with a continuity correction of 0.5). Binomial CIs were calculated by using the Clopper-Pearson interval.

Ethical approval

Ethical approval was not required, as this was a *post hoc* analysis of anonymized participant data from multiple clinical trials, each of which had its own individual ethics approval.

RESULTS

A total of 4 studies were used for this analysis, representing a total of 605 participants (Table I). Risk of bias had been evaluated in a prior systematic review[2]; no study had a high risk of bias or poor external validity. In total, data were available for 592 participants and form the Primary Analysis cohort.

Impact of challenge-stop criteria on baseline challenge thresholds

Fig 2 shows the overall proportion of individuals in the cohort who had peanut allergy and tolerated a given dose at FC. There was a significant difference between all 3 threshold definitions ($P < .001$ [calculated using ANOVA]). This difference was most marked when comparing PRACTALL to the least conservative definition of at least 2 mild CoFAR-grade symptoms: the mean log difference between reaction thresholds was 0.34 (95% CI = 0.29-0.38). The weighted κ statistic was 0.62

TABLE I. Characteristics of the included cohorts

	No. of subjects in each cohort							
		Data included						
Study	**Published**	**Baseline FC**	**Repeat FC**	**Age of cohort**	**Inclusion criteria**	**DBPCFC protocol**	**Challenge-stopping criteria**	**Median cumulative reaction dose at FC**
PEPITES[11]	356	356	109	4-11 y (median 7 y)	Reaction to ≤444 mg	DBPCFC, 30-min intervals (1, 3, 10, 30, 100, and 300 mg of peanut protein)	Based on PRACTALL	144 mg (IQR = 44-444)
BOPI study[12]	64	64	19	8-16 y (median 13 y)	Reaction to ≤4443 mg	DBPCFC, 30- to 60-min intervals (3, 10, 30, 100, 300, 1000, and 3000 mg)	PRACTALL	143 mg (IQR = 43-443)
TRACE study[13]	123	118	71	18-4 y (mean 25 y)	Reaction to ≤1433 mg	DBPCFC, 30- to 60-min intervals (0.003, 0.03, 0.3, 3, 30, 100, 300, and 1000 mg)	Modified PRACTALL	133 mg (IQR = 133-433)
Blumchen et al[14]	62	54	24	3-17 y (median 7 y)	Reaction to ≤4500 mg	Open FC, 2-h intervals (day 1, 3, 10, 30, and 100 mg; day 2, 100, 300, 1000, and 3000 mg; and day 3, 4500 mg)	Objective symptoms according to PRACTALL	143 mg (IQR = 43-1400)

BOPI, Boiled Oral Peanut Immunotherapy; *IQR*, interquartile range; *PEPITES*, Peanut Epicutaneous Immunotherapy Efficacy and Safety; *TRACE*, Threshold Reactivity and Clinical Evaluation. All doses stated are mg peanut protein.

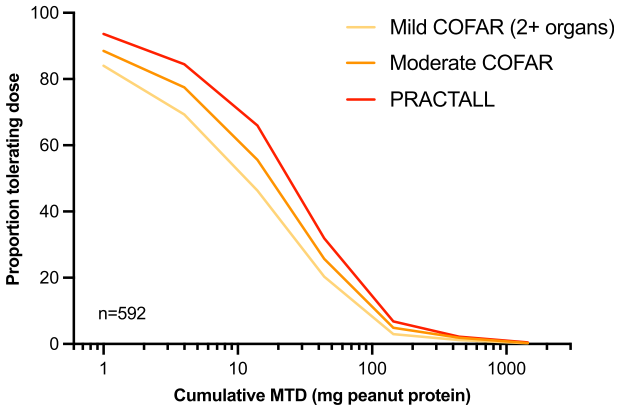


FIG 2. Proportion of individuals in the combined cohort with peanut allergy who tolerated a given dose, with DLS defined according to 1 of 3 definitions: (1) PRACTALL, (2) at least 1 moderate symptom as per CoFAR v3, and (3) a combination of 2 or more mild-grade symptoms in 2 or more systems according to CoFAR v3. *MTD*, Maximum tolerated dose.

(95% CI = 0.57-0.67), indicating a weak-to-moderate level of agreement[15]; discordance between the 2 definitions of DLS was present in 38% of reactions. When PRACTALL was compared with DLS defined according to the presence of at least 1 moderate-grade symptom, the mean log difference was 0.17 (95% CI = 0.14-0.21). The weighted κ statistic was 0.78 (95% CI = 0.74-0.82); discordance was observed in 22% of reactions.

Impact of challenge-stop criteria on reaction severity

The use of less stringent symptoms to define challenge-stop would result in some FCs being terminated earlier, and this might have an impact on symptom severity and the occurrence of anaphylaxis. We therefore evaluated the rate of anaphylaxis (World Allergy Organization 2020 definition[16]) at FC and how (in theory) this might change if the FC were stopped earlier (assuming that no anaphylaxis would occur if the FC were stopped before the dose triggering anaphylaxis). Overall, anaphylaxis occurred in 124 of the FCs (21%) included, although all of the subjects involved were responsive to first-line treatment. When 1 moderate-grade symptom was used as indicating challenge-stop, the rate of anaphylaxis might fall to 17%, and to 14% if the least conservative definition of 2 or more mild CoFAR-grade symptoms was used. However, these differences were not significant by the McNemar test ($P = .5$). Furthermore, the assumption that anaphylaxis would be avoided by stopping earlier may be false: when we used data from Blumchen et al[14] (which used prolonged 2-hour intervals to clarify the symptoms caused by any given FC dose, thus reducing "carryover" of symptoms caused by a preceding dose into the next dosing interval), there was no difference in the rate of anaphylaxis with the different challenge-stop definitions.

Impact of challenge-stop criteria on rate of placebo reactors

In the 3 studies using DBPCFC, we evaluated the rate of symptoms occurring in response to placebo that would have met challenge-stop by the different criteria. We included only those individuals who had the placebo challenge first (before their active challenge), as arguably only the first visit of a DBPCFC is truly blinded in an individual with allergy (if the first visit was "active," in which case the second visit would be assumed to involve placebo and thus not be effectively blinded). The rate

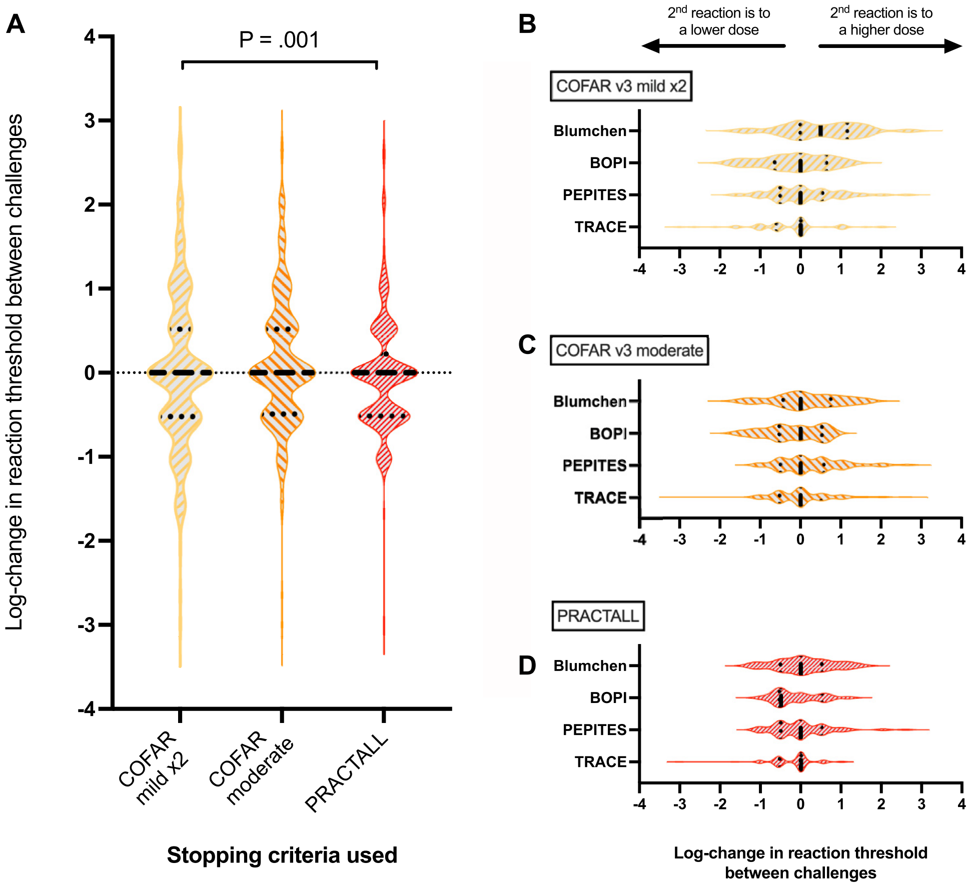


FIG 3. Violin plot of the distributions of the log change in reaction thresholds as a marker of the reproducibility of FCs to peanut. A half-log change in ED is equivalent to a shift in reaction threshold by 1 dosing increment when a PRACTALL-based semi-log regimen is used. Dashed lines represent the median, and dotted lines represent the interquartile range. **A**, Combined study cohort. Change in threshold for each cohort, using the following challenge-stop definitions (*right panel*): a combination of 2 or more mild-grade symptoms in at least 2 systems according to CoFAR v3 (**B**), at least 1 moderate symptom as per CoFAR v3 (**C**); and PRACTALL (**D**). *BOPI,* Boiled Oral Peanut Immunotherapy; *PEPITES*, Peanut Epicutaneous Immunotherapy Efficacy and Safety; *TRACE*, Threshold Reactivity and Clinical Evaluation.

TABLE II. Change in threshold at repeat FC by individual participant data meta-analysis using different challenge-stop criteria (see the text)

	Proportion of participants with	
Criterion	**Maximum half-log change in threshold**	**Maximum 1-log change in threshold**
CoFAR mild × 2	51% (45%-58%)	79% (74%-84%)
≥1 Moderate CoFAR	61% (54%-67%)	87% (82%-91%)
PRACTALL	71% (65%-76%)	95% (90%-96%)
Proposed aligned DLS	67% (61%-73%)	90% (85%-93%)
Reference standard: Patel et al (N = 534)[2]	**71% (56%-83%)**	**91% (84%-95%)**

of placebo reactions (ie, false positives) overall was zero when the PRACTALL criteria were used, but it was 2.9% with moderate COFAR symptoms and 6.6% when the least conservative definition of 2 or more mild CoFAR-grade symptoms was applied.

Impact of challenge-stop criteria on reproducibility of reaction threshold

Overall, 223 participants underwent a subsequent FC in the absence of an intervention (Table I) which might have modified their challenge threshold. The distributions of the log change in reaction thresholds for participants within each included cohort are shown in Fig 3. In terms of the proportion of participants reacting within a particular range, the intraindividual reproducibility was greater with use of PRACTALL than with either CoFAR v3 criteria for DLS ($P < .001$) (Table II).

Impact on efficacy outcomes in OIT studies

We also evaluated the impact of the 3 different challenge-stop criteria on the change in reaction threshold following active OIT in the 2 relevant studies of OIT (combined sample size n = 66).[12,14] The mean log fold increase in reaction threshold with use of the least conservative definition of 2 or more mild CoFAR-grade symptoms was 1.83 (95% CI = 1.6-2.0), which was significantly higher than that seen with the other 2 definitions.

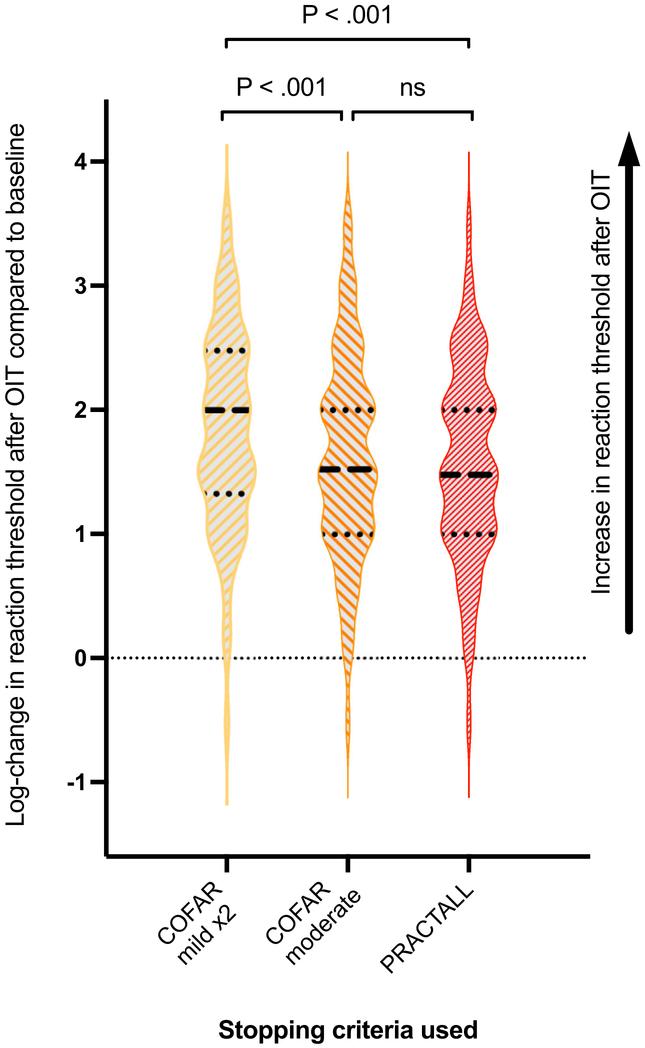


FIG 4. Violin plot of the distribution of the log change in reaction thresholds following peanut OIT in the 2 included studies[12,14] defined according to the 3 challenge-stop criteria assessed. Dashed lines represent the median, and the dotted lines represent the interquartile range.

There was no significant difference seen when we compared DLS based on at least 1 moderate or severe CoFAR symptom (1.61 [95% CI = 1.40-1.82]) with the PRACTALL criteria (1.56 [95% CI = 1.36-1.75]) (Fig 4). Before OIT, it is not uncommon to see mild symptoms before the dose causing DLS; however, following OIT, these low-grade symptoms tend not to occur (Fig 5[17]).

Data-driven modification of CoFAR v3 challenge-stop criteria to generate a new framework for challenge stop criteria

Finally, we used a data-driven approach to assess the impact of modifying the CoFAR v3 framework for defining challenge-stop—the objective being to mitigate against the impact of including less objective symptoms as DLS (with PRACTALL used as the reference standard). The optimal set of modifications to CoFAR v3 (defined as those having the least adverse impact compared with PRACTALL) are presented in Table I and include the following: (1) redefining the significance of mild subjective or minimal cutaneous signs to be aligned with PRACTALL/Learning Early about Peanut Allergy (LEAP) study criteria,[8] (2) classifying mild rhinitis symptoms as never dose limiting; (3) including ocular symptoms as per PRACTALL; (4) requiring the presence of at least 1 other symptom from a different system when including throat tightness as a DLS, (5) considering nausea as being distinct to abdominal pain/discomfort in terms of relevance as a DLS, and (6) removing the possibility of a single episode of emesis contributing to DLS when due to a gag reflex and including a single episode of emesis as a DLS when it is considered to be due to an allergic reaction (in contrast to PRACTALL).

Using the proposed aligned DLS as shown in Table I, there was much less discordance between the new aligned DLS and PRACTALL (Fig 6, *A*), with a weighted κ statistic of 0.89 (95% CI = 0.87-0.91), indicating concordance in 89% of reactions. Reproducibility was also favorable compared to PRACTALL (Fig 6, *B*), with significantly less difference than with CoFAR v3 and minimal difference versus with the reference standard (Table II). When the proposed aligned DLS were used, there was no significant difference in the rate of possible anaphylaxis (18.2% vs 20.9% with PRACTALL [$P = .5$ according to the McNemar test]). Finally, when the new aligned DLS definition was used to reassess efficacy in the 2 OIT studies, there was no difference in outcome between the use of moderate-grade COFAR symptoms and use of PRACTALL (Fig 6, *C*).

DISCUSSION

In this secondary analysis of FC data from 4 clinical trials, we demonstrated that including less objective symptoms as challenge-stop criteria adversely affects the validity of FC as a tool to measure clinical efficacy in food allergy intervention studies. Applying CoFAR DLS definitions instead of PRACTALL resulted in a relative underestimate of the challenge threshold, equivalent to almost a single dosing increment when using a semi-log FC protocol as recommended by PRACTALL. The use of less objective criteria also adversely affected both the rate of placebo reactions that would have met stopping criteria, as well as reproducibility of FC threshold, without significantly reducing the rate of reported anaphylaxis. Finally, applying the least conservative interpretation of CoFAR v3 DLS (2 mild symptoms occurring in ≥2 organs) to assess efficacy outcomes from the 2 OIT studies resulted in a significant 15% overestimate of the change in reaction thresholds. This implies that the use of more significant subjective or objective symptoms is required to avoid compromising the utility of FCs for evaluation of changes in clinical reactivity following therapeutic intervention.

The PRACTALL consensus report on standardizing DBPCFC was generated following discussions in 2008 with the aim to develop an international standard to facilitate comparisons between "studies involving diagnosis, natural history, and therapeutic trials in food allergy" undertaken globally.[1] Although the document states that "there are currently no agreed upon published parameters [to define challenge-stop] because clinical judgment is needed, and circumstances might vary by patient or study characteristics," a key part of the report was designed to provide a framework to inform challenge-stop criteria, typically on the basis of at least 1 significant objective symptom, or a combination of subjective and/or mild objective symptoms or signs occurring in more than 1 organ system. Many studies have used

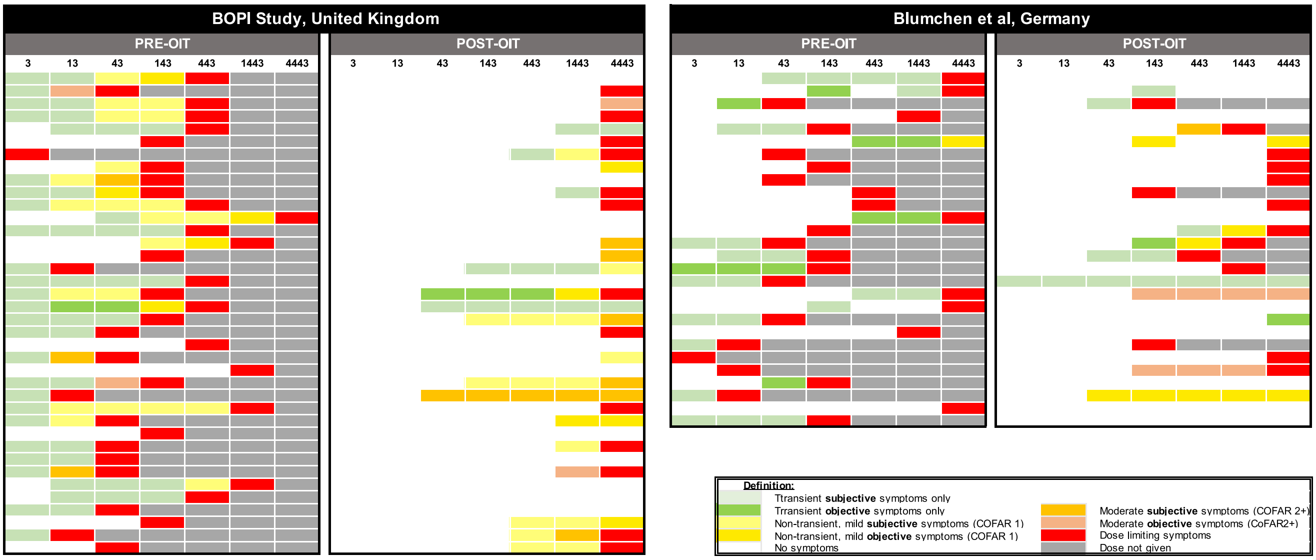


FIG 5. Heatmap demonstrating the type of symptoms experienced at FC before (PRE-) and following (POST-) OIT in the 2 included studies. Transient symptoms defined as CoFAR grade 1 symptoms that resolve within 1 challenge dosing interval. CoFAR grade symptoms defined according to CoFAR grading system, version 2.[17] *BOPI,* Boiled Oral Peanut Immunotherapy.

this framework as a basis for defining DLS in study protocols. However, more recent studies (including some phase 3 commercial studies) have included less objective symptoms, sometimes from just a single organ system, as DLS. However, the inclusion of such symptoms, particularly subjective oropharyngeal symptoms, is associated with a high rate of false-positive challenges.[4] Chinthrajah et al recently reported that CoFAR has developed guidance for determining DLS at FC in an attempt to reduce the heterogeneity seen in challenge-stopping criteria across clinical trials.[10] As can be seen in Fig 1, there are some significant areas of divergence from both PRACTALL and the challenge-stop criteria used in the LEAP study,[8] which have also been used as a basis for DLS in some studies.[9] Under CoFAR v3, mild erythema in isolation would be considered a DLS, whereas occasional itching in combination with another subjective symptom might trigger challenge-stop. Throat tightness would also be considered a DLS even if not severe. A single episode of vomiting due to gagging (for example, as a result of taste aversion) would be considered a DLS if occurring in combination with another subjective symptom. Similarly, persistent moderate nausea with a decreased level of activity would also trigger challenge-stop.[10] This would also create diagnostic difficulties if these DLS were extrapolated to routine clinical practice outside the clinical trials setting.

We have previously highlighted our concern that some of these symptoms (eg, throat tightness without hoarseness in isolation [which in our experience, is often transient in the first instance and can also occur during placebo FC], more than a few areas of erythema [but not widespread erythema], or "pruritus causing protracted scratching" [which is common in children with eczema under stressful conditions]) might be sufficient to terminate a challenge.[18] The analysis in this report provides objective evidence that including these symptoms as DLS can significantly underestimate the challenge threshold by around 1 dosing increment. Furthermore, although the aim might be to reduce the rate of anaphylaxis at FC, we did not find any statistically significant reduction in anaphylaxis at FC when these less objective stopping criteria were used, which is a finding consistent with a previous analysis of 652 open FCs that was reported by Nachshon et al.[4] Of concern, we did find a higher rate of placebo reactions that would meet the less objective stopping criteria, particularly with the least conservative criterion of 2 or more mild CoFAR-grade symptoms. Furthermore, the analysis in Fig 5 confirms that these subjective symptoms often respond very well to food allergy desensitization, so there is a risk that categorizing such subjective symptoms as dose limiting could underestimate an individual's reaction threshold before immunotherapy and thus exaggerate the true impact of the intervention. This was less of an issue when moderate-grade DLS were used.

A key observation was the impact of less objective DLS on the reproducibility of FC thresholds. We recently reported that DBPCFCs, when conducted according to the PRACTALL consensus criteria, are a relatively reproducible tool to assess efficacy outcomes in clinical trials. Our analysis found that in most individuals with allergy to peanut or cow's milk (70%-80%), the ''shift'' in threshold over time was limited to a half-log, which is equivalent to a single dosing increment when using a semi-log PRACTALL dosing regimen (eg, a change in reaction threshold from 30 mg to 100 mg of protein).[2,3] When non-PRACTALL criteria are used, the reproducibility of reaction thresholds is compromised, particularly when less objective, mild symptoms in combination are used as DLS.

Nonetheless, the application of PRACTALL stopping criteria can be problematic. It is not unusual for individuals with allergy to experience predominantly gastrointestinal symptoms at FC.[19,20] Under a "purist" interpretation of PRACTALL, abdominal pain and a single episode of vomiting would constitute only 2 "orange" symptoms and not the 3 often used to imply a reaction. That being said, PRACTALL does not require 3 orange symptoms but instead allows "clinical judgment" in assessing the relevance of such symptoms (although in practice, PRACTALL is often implemented as needing 3 orange symptoms or 1 red symptom to

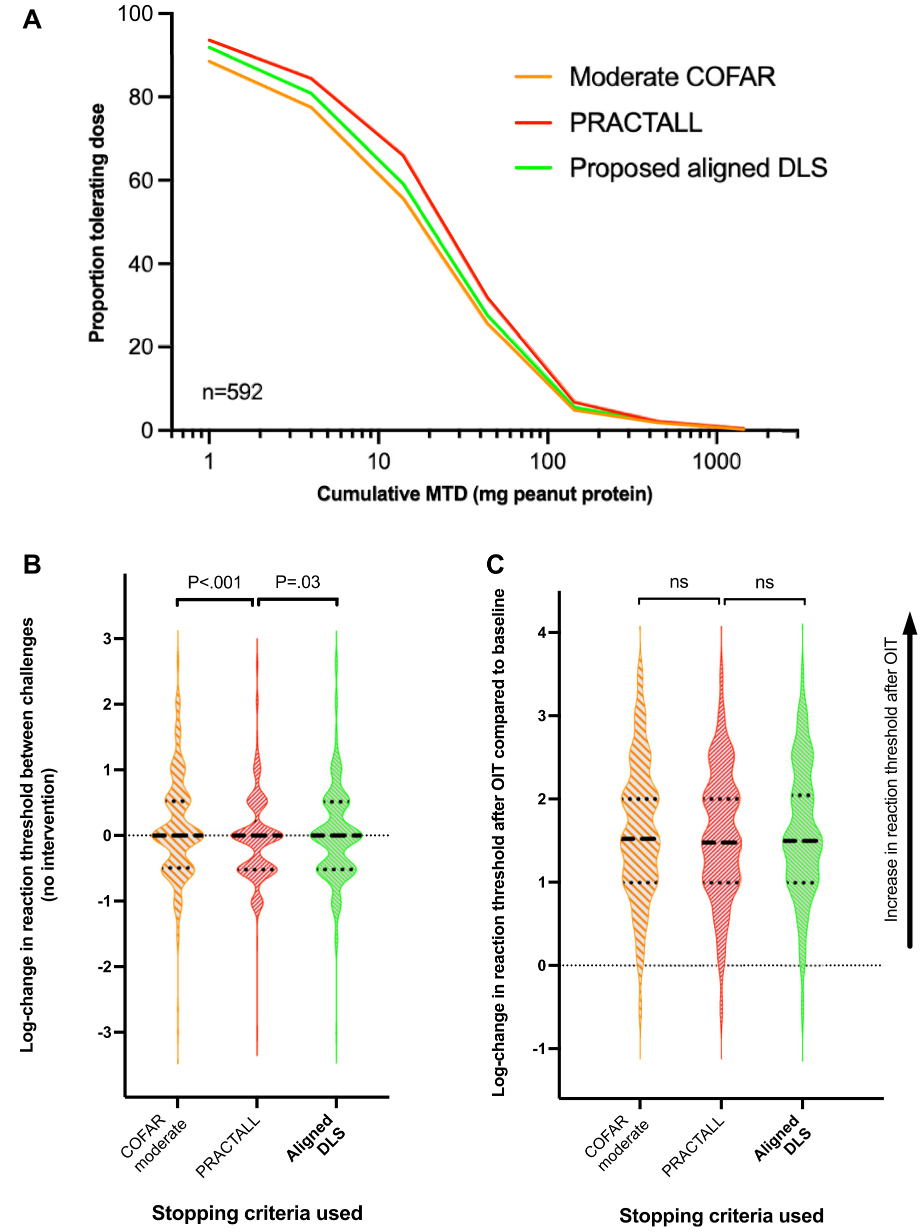


FIG 6. Evaluation of the proposed aligned DLS (Table I) compared with moderate-grade CoFAR symptoms and PRACTALL: at baseline peanut challenge **(A)**, in terms of reproducibility of challenge threshold in those individuals who underwent a subsequent FC without therapeutic intervention **(B)**, and when assessing the increase in challenge threshold in participants who underwent OIT (**C**).

define challenge-stop). However, this raises the question of how much discretion can be allowed in the context of clinical trials in which a more rigid, protocol-driven approach is required. In comparing PRACTALL with CoFAR v3 (Fig 1), perhaps the main differences can be summarized as follows: (1) CoFAR v3 permits any degree of erythema beyond a few faint areas or protracted scratching as a DLS; (2) throat tightness is a DLS in CoFAR v3 that can be problematic, as it is a very subjective symptom and is relatively common during FCs, occurring in 16% of challenges included in this analysis before PRACTALL stopping criteria have been achieved; (3) CoFAR v3 includes an "isolated emesis thought to be secondary to gag" as a mild symptom (arguably, such an occurrence is not an allergic event at all but a consequence of taste aversion); and (4) COFAR also allows for "persistent nausea with decreased activity" as a DLS, even though this too can be a consequence of taste aversion.

On the basis of the aforementioned, to avoid the various issues with both PRACTALL and the use of mild CoFAR symptoms as DLS, we used a data-driven approach to test various modifications of the CoFAR v3 framework. The optimal framework, shown in Fig 1, seeks to align PRACTALL with CoFAR v3 and includes moderate abdominal discomfort with reduced activity as a DLS. Oropharyngeal pruritus is no longer a potential DLS; however, when there are no clear objective "red" symptoms present, only 2 concurrent DLS (from different organ systems) are required to achieve challenge-stop. Our hope is that this new approach to defining DLS will promote discussion to achieve a broader and inclusive consensus of which challenge-stopping criteria should be used in future clinical trials.

Although this analysis includes FC data from only 592 participants, the sample remains larger than that in any clinical trial reported to date. Furthermore, in terms of reaction thresholds and extent of reproducibility, the cohort in this analysis was similar to that our previous report of more than 3000 DBPCFC reported for peanut, an analysis that was robust and included multiple sensitivity analyses.[2] Whereas 3 of the included studies utilized DBPCFC, the fourth (Blumchen et al[14]) used open challenge methodology, although otherwise conducted according to PRACTALL. Double-blind methodology was not feasible in this study because the challenge protocol used a prolonged 2-hour dosing interval to specifically evaluate the duration of symptoms (which would otherwise be obscured with shorter intervals). We did perform a sensitivity analysis comparing the studies utilizing DBPCFC and the study by Blumchen et al,[14] and we did not observe any differences (data not shown). Importantly, inclusion of the study by Blumchen et al[14] also allowed a more realistic evaluation of any reduction in rate of anaphylaxis when using less objective symptoms because the prolonged challenge intervals reduced “carryover” of symptoms into the next challenge interval. We did not observe any difference in the rate of anaphylaxis when using less conservative criteria to determine challenge-stop, which thus challenges the assumption that stopping FC at an average of only 1 dosing interval results in a significant reduction in anaphylaxis at FC.

The assessment of reproducibility might be limited by the inclusion of 3 pediatric studies, which by definition included children who might be more likely to outgrow their peanut allergy. This is reflected in the slight skewing of the “tail” in Fig 3, *D*, which likely reflects natural tolerance within some subjects allocated to the placebo arm in the included studies; even in these studies, however, the shift in threshold was no greater than that seen in the Threshold Reactivity and Clinical Evaluation (TRACE) study, which was performed in adults with peanut allergy: in the latter, 68% of participants had a shift in threshold limited to a half-log change, compared with the 67% to 74% of children in the 3 pediatric studies. Although ideally this analysis would have included additional cohorts, we were unable to identify other cohorts for which sufficient detailed challenge and symptom data (relating to both objective and subjective symptoms) had been recorded and could be provided under data protection legislation. Of note, the majority of clinical trials evaluated for potential inclusion in this analysis omitted to provide sufficient detail regarding challenge-stop criteria in their publications.

Notwithstanding, our analysis is, to our knowledge, the first to compare the performance of the new proposed CoFAR v3 challenge-stop criteria against PRACTALL by using real-world data from clinical trials. We have demonstrated that the inclusion of less objective symptoms as DLS, particularly those included in the mild CoFAR v3 proposal, may compromise the validity of the FC as a tool with which to assess efficacy. This can be mitigated against by excluding such mild symptoms and, in the absence of a clear objective DLS, requiring the presence of significant (subjective) symptoms from at least 2 organ systems. We have presented 1 such option for relevant stakeholders to consider in the hope that this will stimulate a wider discussion and result in a new inclusive consensus for DLS to be implemented in the context of clinical trials.

We thank Lea Blum, Beatriz González Cano, and Klara Liddell for their contribution in cleaning the raw data used in this analysis. We are grateful to DBV Technologies and the UK Food Standards Agency for providing data from the relevant clinical trials for inclusion in this work.

Clinical implications: Using less objective symptoms to define challenge-stop in FCs underestimates the challenge threshold and reduces the validity of FCs as a tool to measure efficacy outcomes in interventional studies.

REFERENCES

1. Sampson HA, Gerth van Wijk R, Bindslev-Jensen C, Sicherer S, Teuber SS, Burks AW, 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;130:1260-74.
2. Patel N, Adelman DC, Anagnostou K, Baumert JL, Blom WM, Campbell DE, et al. Using data from food challenges to inform management of consumers with food allergy: a systematic review with individual participant data meta-analysis. J Allergy Clin Immunol 2021;147:2249-62.
3. Turner PJ, Patel N, Campbell DE, Sampson HA, Maeda M, Katsunuma T, et al. Reproducibility of food challenge to cow’s milk: systematic review with individual participant data meta-analysis. J Allergy Clin Immunol 2022;150:1135-43.e8.
4. Nachshon L, Zipper O, Levy MB, Goldberg MR, Epstein-Rigby N, Elizur A. Subjective oral symptoms are insufficient predictors of a positive oral food challenge. Pediatr Allergy Immunol 2021;32:342-8.
5. Rodríguez Del Río P, Escudero C, Sánchez-García S, Ibáñez MD, Vickery BP. Evaluating primary end points in peanut immunotherapy clinical trials. J Allergy Clin Immunol 2019;143:494-506.
6. PALISADE Group of Clinical Investigators, Vickery BP, Vereda A, Casale TB, Beyer K, du Toit G, et al. AR101 Oral immunotherapy for peanut allergy. N Engl J Med 2018;379:1991-2001.
7. O’B Hourihane J, Beyer K, Abbas A, Fernández-Rivas M, Turner PJ, Blumchen K, et al. Efficacy and safety of oral immunotherapy with AR101 in European children with a peanut allergy (ARTEMIS): a multicentre, double-blind, randomised, placebo-controlled phase 3 trial. Lancet Child Adolesc Health 2020;4:728-39.
8. Du Toit G, Roberts G, Sayre PH, Bahnson HT, Radulovic S, Santos AF, et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. N Engl J Med 2015;372:803-13.
9. Jones SM, Kim EH, Nadeau KC, Nowak-Wegrzyn A, Wood RA, Sampson HA, et al. Efficacy and safety of oral immunotherapy in children aged 1-3 years with peanut allergy (the Immune Tolerance Network IMPACT trial): a randomised placebo-controlled study. Lancet 2022;399:359-71.
10. Chinthrajah RS, Jones SM, Kim EH, Sicherer SH, Shreffler W, Lanser BJ, et al. Updating the CoFAR grading scale for systemic allergic reactions in food allergy. J Allergy Clin Immunol 2022;149:2166-70.e1.
11. Fleischer DM, Greenhawt M, Sussman G, Begin P, Nowak-Wegrzyn A, Petroni D, et al. Effect of epicutaneous immunotherapy vs placebo on reaction to peanut protein ingestion among children with peanut allergy: the PEPITES randomized clinical trial. JAMA 2019;321:946-55.
12. Patel N, Vazquez-Ortiz M, Lindsley S, Campbell DE, Turner PJ. Low frequency of soya allergy in peanut-allergic children: relevance to allergen labelling on medicines. Allergy 2018;73:1348-50.
13. Dua S, Ruiz-Garcia M, Bond S, Durham SR, Kimber I, Mills C, et al. Effect of sleep deprivation and exercise on reaction threshold in adults with peanut allergy: a randomized controlled study. J Allergy Clin Immunol 2019;144:1584-94.e2.
14. Blumchen K, Trendelenburg V, Ahrens F, Gruebl A, Hamelmann E, Hansen G, et al. Efficacy, safety, and quality of life in a multicenter, randomized, placebo-controlled trial of low-dose peanut oral immunotherapy in children with peanut allergy. J Allergy Clin Immunol Pract 2019;7:479-91.
15. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
16. Cardona V, Ansotegui IJ, Ebisawa M, El-Gamal Y, Fernandez Rivas M, Fineman S, et al. World Allergy Organization anaphylaxis guidance 2020. World Allergy Organ J 2020;13:100472.
17. Burks AW, Jones SM, Wood RA, Fleischer DM, Sicherer SH, Lindblad RW, et al, Consortium of Food Allergy Research (CoFAR). Oral immunotherapy for treatment of egg allergy in children. N Engl J Med 2012;367:233-43.
18. Turner PJ, Patel N, Mäkelä MJ, Kukkonen K, Deschildre A, Blumchen K, et al. Improving the reporting of allergic adverse events during immunotherapy for food allergy. J Allergy Clin Immunol 2022;150:1242-4.

19. Frugier C, Graham F, Samaan K, Paradis L, Des Roches A, Bégin P. Potential Efficacy of high-dose inhaled salbutamol for the treatment of abdominal pain during oral food challenge. J Allergy Clin Immunol Pract 2021;9:3130-7.

20. Dua S, Ruiz-Garcia M, Bond S, Dowey J, Durham SR, Kimber I, et al. Effects of exercise and sleep deprivation on reaction severity during oral peanut challenge: a randomized controlled trial. J Allergy Clin Immunol Pract 2022;10:2404-13.e1.