

Journal Pre-proof



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

Paul J. Turner, FRCPCH PhD, Nandinee Patel, MRCPCH PhD, Katharina Blumchen, MD, Stefanie Berkes, BSc, Hugh A. Sampson, MD, Kirsten Beyer, MD

PII: S0091-6749(22)02591-X

DOI: <https://doi.org/10.1016/j.jaci.2022.12.818>

Reference: YMAI 15818

To appear in: *Journal of Allergy and Clinical Immunology*

Received Date: 21 August 2022

Revised Date: 29 November 2022

Accepted Date: 20 December 2022

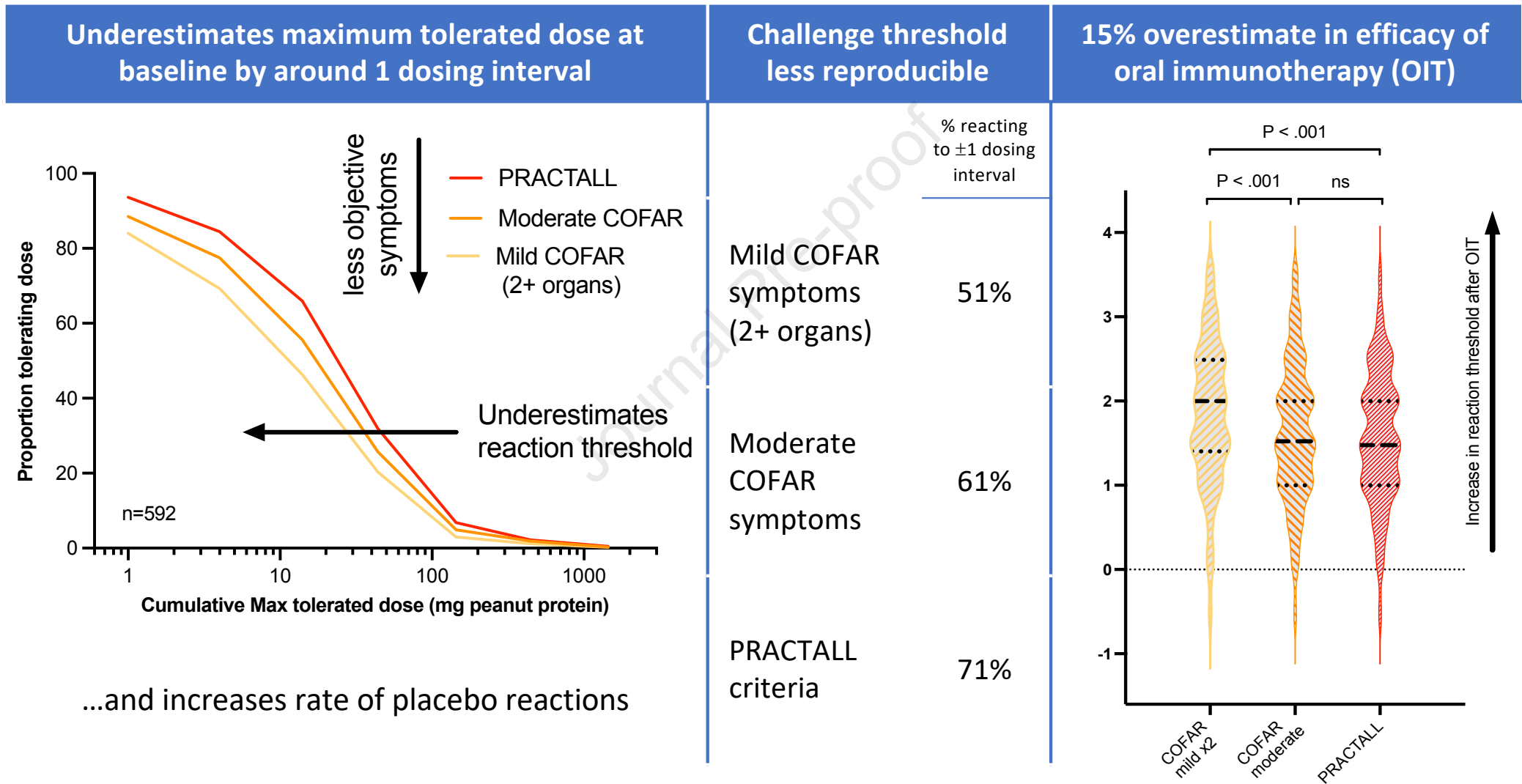
Please cite this article as: Turner PJ, Patel N, Blumchen K, Berkes S, Sampson HA, Beyer K, Impact of using less objective symptoms to define tolerated dose during food challenges: a data-driven approach, *Journal of Allergy and Clinical Immunology* (2023), doi: <https://doi.org/10.1016/j.jaci.2022.12.818>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Inc. on behalf of the American Academy of Allergy, Asthma & Immunology.



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



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, Kirsten Beyer MD^d

Affiliation:

^aNational Heart & Lung Institute, Imperial College London, London, United Kingdom;

^bDepartment of Paediatric and Adolescent Medicine, Pneumology, Allergology and Cystic fibrosis, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany;

^cDivision of Pediatric Allergy and Immunology, Jaffe Food Allergy Institute, Icahn School of Medicine at Mount Sinai, New York, USA;

^dDepartment of Pediatric Respiratory Medicine, Immunology and Critical Care Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany.

***Corresponding author:**

Dr Paul Turner

National Heart & Lung Institute,

Imperial College London,

Norfolk Place

London, W2 1PG

Email: p.turner@imperial.ac.uk

Tel: +44 20 3312 7754

Funding: Supported by a UK Medical Research Council award to P.J. Turner (ref MR/S036954/1).

Declaration of interests

P.J. Turner reports grants from UK Medical Research Council, NIHR/Imperial BRC and JM Charitable Foundation; personal fees from UK Food Standards Agency, Aimmune Therapeutics, Allergenics, 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, 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 (NIH)/National Institute of Allergy and Infectious Diseases (NIAID), 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, Nestle; speakers bureau for Aimmune Therapeutics, ALK, Danone, Infectopharm, Med Update, Nestle, Nutricia; and research grants from Aimmune, ALK, Danone, DBV, Hipp, Hycor, Infectopharm, Nutricia, outside the submitted work. The other authors have nothing to disclose.

ABSTRACT:

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

Objective: 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: Analysis of individual participant data (IPD) including IPD meta-analysis, using three different published challenge-stop criteria: (i) PRACTALL; (ii) CoFAR v3 with at least 1 moderate/severe grade symptom; or (iii) CoFAR v3 with at least 2 mild symptoms occurring in different organ systems. Reproducibility of challenge threshold was also assessed in participants undergoing subsequent repeat FC.

Results: 4 studies, with detailed challenge data from a total of 592 participants, were included. Applying CoFAR v3 definitions for dose-limiting symptoms (DLS) resulted in an underestimate of reaction thresholds compared to PRACTALL ($P < 0.001$), equivalent to almost a single dosing increment when using a semi-log dosing regimen. Reproducibility was also reduced when applying CoFAR v3 ($P < 0.001$, $N = 223$). Using the least conservative interpretation of CoFAR v3 (≥ 2 mild symptoms occurring in different systems) resulted in a significant overestimate of 15% when assessing OIT efficacy. Applying a data-driven, minor modification to CoFAR v3 resulted in a new set of challenge-stop criteria with similar validity to PRACTALL, but simpler in implementation and one where significant gastrointestinal discomfort with observable decreased activity remains a DLS.

Conclusion: The use of less objective symptoms to define challenge-stop compromises the reproducibility of FC as a tool to assess efficacy outcomes in interventional studies, and potentially over-estimates the efficacy of the intervention tested.

Clinical Implication:

Using less-objective symptoms to define challenge-stop in food challenges underestimates the challenge threshold, and reduces the validity of food challenges as a tool to measure efficacy outcomes in interventional studies.

Capsule Summary

Using a data-driven approach, we propose a modification of the CoFAR v3 challenge-stop criteria. This mitigates against the adverse impact of including less objective symptoms, while providing a simpler framework for challenge stop, which includes moderate-grade gastrointestinal symptoms.

Keywords

CoFAR, Eliciting dose, Food challenge, Peanut, PRACTALL, Reproducibility, Thresholds.

90 **Abbreviations:**

91	95%CI	95% confidence interval
92	AIT	Allergy immunotherapy
93	CoFAR	Consortium of Food Allergy Research
94	DBPCFC	Double-blind placebo-controlled food challenge
95	DLS	Dose limiting symptoms
96	FC	Food challenge
97	IPD	Individual participant data
98	MTD	Maximum tolerated dose
99	OIT	Oral immunotherapy

100

INTRODUCTION

Oral food challenges (FC) are a key tool to evaluate the impact of interventions which might alter clinical reactivity in food allergy. While oral immunotherapy (OIT) is becoming an increasingly mainstream intervention, there is still a need for further research to evaluate protocols which 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 (AIT).

The gold-standard method for FC is the double-blind, placebo-controlled food challenge (DBPCFC). DBPCFC are frequently used to determine the primary outcome in clinical trials; therefore, robust and standardized protocols are needed. The AAAAI-EAACI PRACTALL consensus report on standardizing DBPCFC, published in 2012,¹ quickly gained international acceptance and has been widely used for over a decade to inform the execution of DBPCFC including challenge-stop criteria. More recently, we evaluated the reproducibility of DBPCFC to peanut and cow's milk, and found that intra-individual 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, equivalent to a single dosing increment when using a semi-log dosing regimen as recommended by PRACTALL.^{2,3}

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

However, not all clinical trials have used PRACTALL consensus criteria to determine challenge-stop.⁵ Some studies have utilized less rigid criteria – often referred to as dose-

126 limiting symptoms (DLS) – for this purpose.^{6,7} Conversely, some studies have used
127 alternatives which better define challenge-stop symptoms (compared to PRACTALL).^{8,9}
128 Chinthrajah et al. recently proposed an updated framework (reproduced in Figure 1), as
129 part of the updated Consortium for Food Allergy Research (CoFAR) Grading Scale for
130 Systemic Allergic Reactions.¹⁰ This includes the possibility of non-objective, subjective
131 symptoms in a single organ system being considered dose-limiting and leading to stopping
132 a FC. For example, under the new proposal, isolated symptoms such as “throat tightness
133 without hoarseness” or “more than a few areas of erythema” (but not widespread erythema)
134 are sufficient to terminate a challenge.

135 In this analysis, we re-evaluated detailed challenge data available from clinical trials in
136 peanut-allergic participants, applying both PRACTALL and the new CoFAR version 3
137 criteria, to assess the impact of using less objective challenge-stop criteria.

METHODS

Study selection

Having previously undertaken a systematic review of the literature to identify studies which used DBPCFC to evaluate clinical reaction thresholds to peanut, we included 3 studies where detailed symptom data (relating to both subjective and objective symptoms at each FC dose) were available and individuals had undergone a further FC following initial DBPCFC (conducted according to the same protocol). Two were AIT studies,^{11,12} while the third was a study in which peanut-allergic adults with positive DBPCFC were randomized to undergo repeat FC with or without co-factors (exercise, sleep deprivation).¹³ For the purpose of this analysis, data were used from the baseline DBPCFC and non-intervention challenge (without co-factor) 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, but instead utilized an open FC protocol with a dosing interval of 2 hours, but which were 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, using the following 3 definitions:

- i. the dose causing challenge-stopping symptoms according to the PRACTALL consensus;
- ii. the lowest dose associated with the occurrence of at least one “moderate” (or “severe”) grade symptom defined according to the proposed CoFAR v3 criteria (see Figure 1)

- iii. the lowest dose associated with the occurrence of at least two “mild” symptoms according to CoFAR v3 occurring in >1 different organ system.

For all cohorts, the cumulative maximum tolerated dose (MTD) was set as the dose given immediately prior to that causing DLS. Any discrepancies identified 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) and R software, (version 4.0.3). All statistical tests were two-tailed, and a P-value of <.05 was considered significant. Agreement was assessed using kappa with linear weighting.¹⁵

To assess the reproducibility of challenge thresholds over time within individuals, we extracted individual participant data of 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 assessed using Violin plots. A similar process was followed to evaluate log-fold change in reaction threshold in participants who underwent OIT in 2 of the 4 included trials.^{12,14}

Finally, we then evaluated the impact 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 to PRACTALL criteria. Normality of distribution was assessed using the D'Agostino-Pearson test, and the data then used for IPD-meta-analysis. Rates were pooled across studies 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 confidence intervals were calculated using the Clopper–Pearson interval.

188

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 their own individual ethics approval.

193

RESULTS:

Four 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.

199

Impact of challenge-stop criteria on baseline challenge thresholds

Figure 2 shows the overall proportion of peanut-allergic individuals in the cohort tolerating a given dose at FC. There was a significant difference between all 3 threshold definitions ($P < 0.001$, ANOVA). This was most marked when comparing PRACTALL to the least conservative definition of ≥ 2 “mild” CoFAR-grade symptoms: the mean log-difference between reaction thresholds was 0.34 (95%CI 0.29-0.38). Weighted kappa was 0.62 (95%CI 0.57-0.67), indicating a “weak-moderate” level of agreement;¹⁵ discordance between the 2 definitions of DLS was present in 38% of reactions. Comparing PRACTALL to DLS defined according to the presence of at least one “moderate” grade symptom, the mean log-difference was 0.17 (95%CI 0.14-0.21). Weighted kappa of 0.78 (95%CI 0.74-0.82); discordance was observed in 22% of reactions.

211

Impact of challenge-stop criteria on reaction severity

The use of less stringent symptoms to define challenge-stop would result in some FC being terminated earlier, and this might impact on symptom severity and the occurrence of anaphylaxis. We therefore evaluated the rate of anaphylaxis (World Allergy Organization 2020 definition¹⁶) at FC, and how (in theory) this might change if the FC had been stopped earlier (assuming that no anaphylaxis would occur if the FC was stopped before the dose triggering anaphylaxis). Overall, anaphylaxis occurred in 124 (21%) of FC included, although all were responsive to first-line treatment. When using one “moderate” grade symptom as indicating challenge-stop, the rate of anaphylaxis might fall to 17%, and to 14% if using the least conservative definition of ≥ 2 “mild” CoFAR-grade symptoms. However, these differences were not significant by McNemar’s test ($P=0.5$). Furthermore, the assumption that anaphylaxis would be avoided by stopping earlier may be false: when analysing data from Blumchen et al (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 methodology, we evaluated the rate of symptoms occurring to placebo which would have met challenge-stop by the different criteria. We included only 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 allergic individual (if the first visit was “active”, then the second visit would be assumed to be placebo and thus not effectively blinded). The rate of placebo reactions (i.e. false positives) overall was zero when using

PRACTALL criteria, but 2.9% with moderate COFAR symptoms and 6.6% when applying the least conservative definition of ≥ 2 “mild” CoFAR-grade symptoms.

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 Figure 3. Intra-individual reproducibility was greater when using PRACTALL compared to either CoFAR v3 criteria for DLS ($P < 0.001$), in terms of the proportion of participants reacting within a particular range (Table II).

Impact on efficacy outcomes reported to oral immunotherapy (OIT)

We also evaluated the impact of the 3 different challenge-stop criteria on the change in reaction threshold following active OIT in the two relevant studies of OIT (combined sample size, $N=66$).^{12,14} The mean log-fold increase in reaction threshold using the least conservative definition of 2+ “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. There was no significant difference seen when comparing DLS on the basis of at least 1 “moderate”/“severe” CoFAR symptom (1.61, 95%CI 1.40-1.82) to the PRACTALL criteria (1.56, 95%CI 1.36-1.75) (Figure 4). Prior to OIT, it is not uncommon to see mild symptoms prior to the dose causing DLS; however, following OIT, these low-grade symptoms tend not to occur (Figure 5).

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 (using PRACTALL as the reference standard). The optimal set of modifications to CoFAR v3 (defined as those having least adverse impact compared to PRACTALL) are presented in the Table I, and include:

- Redefining the significance of mild subjective or minimal cutaneous signs to be aligned with PRACTALL/LEAP study criteria⁸
- Classifying mild rhinitic symptoms as never dose-limiting.
- Including ocular symptoms as per PRACTALL
- Requiring at least one other symptom from a different system to be present when including throat tightness as a DLS.
- Consideration of nausea as being distinct to abdominal pain/discomfort in terms of relevance as a DLS
- 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 where this 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 (Figure 6A) with a weighted kappa of 0.89 (95%CI 0.87-0.91), indicating concordance in 89% of reactions. Reproducibility was also favourable compared to PRACTALL (Figure 6B), with significantly less difference compared to CoFAR v3 and minimal difference to the reference standard (Table II). Using the

proposed aligned DLS, there was no significant difference in the rate of possible anaphylaxis (18.2% compared to 20.9% with PRACTALL, $P=0.5$, McNemar's test). Finally, when the new aligned DLS definition was used to re-assess efficacy in the two OIT studies, there was no difference in outcome between the use of moderate-grade COFAR symptoms or PRACTALL (Figure 6C).

DISCUSSION

In this secondary analysis of FC data from 4 clinical trials, we demonstrated that including less objective symptoms as challenge-stop criteria adversely impacts upon 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 impacted both the rate of placebo reactions which 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 over-estimate of the change in reaction thresholds of 15%. This implies that the use of more significant subjective or objective symptoms is required to avoid compromising the utility of FC to evaluate 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 While the document states “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 to provide a framework to inform challenge-stop criteria, typically on the basis of at least one 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 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 food challenge, in an attempt to reduce the heterogeneity seen in challenge-stopping criteria across clinical trials.⁷ As can be seen in Figure 1, there are some significant areas of divergence from both PRACTALL and also 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, while 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 vomit 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, for example “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” (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, while 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 using these less-objective stopping criteria, a finding consistent with a previous analysis of 652 open FC reported by Nachshon et al.⁴ Of concern, we did find a higher rate of placebo reactions which would meet the less-objective stopping criteria, particularly with least conservative criteria of ≥ 2 “mild” CoFAR-grade symptoms. Furthermore, the analysis in Figure 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 pre-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 DBPCFC, when conducted according to PRACTALL consensus criteria, are a relatively reproducible tool to assess efficacy outcomes in clinical trials. Our analysis found in most (70-80%) individuals with food allergy to peanut or cow’s milk, the “shift” in threshold over time was limited to a half-log, equivalent to a single dosing increment when using a semi-log PRACTALL dosing regimen – for example, 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 allergic individuals to experience predominantly gastrointestinal symptoms at FC.^{19,20} Under a “purist” interpretation of PRACTALL, abdominal pain and a single episode of vomiting would only constitute 2 “orange” symptoms and not the 3 often used to imply a reaction. That being said, PRACTALL does not require 3 “orange” symptoms but rather, allows “clinical judgement” in assessing the relevance of such symptoms (although in practice, PRACTALL is often implemented as needing 3 “orange” symptoms). This raises the question over how much discretion can be allowed, however, in the context of clinical trials where a more rigid, protocol-driven approach is required. In comparing PRACTALL to CoFAR v3 (Figure 1), perhaps the main differences can be summarized as follows:

- CoFAR v3 permits any degree of erythema beyond a few faint areas, or protracted scratching as a DLS
- Throat tightness is a DLS in CoFAR v3; this can be problematic, as this is a very subjective symptom and is relatively common during FC, occurring in 16% of challenges included in this analysis prior to PRACTALL stopping criteria being achieved
- 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
- COFAR also allows for “persistent nausea with decreased activity” as a DLS, even though this too can be a consequence of taste aversion.

382

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

392

393 While this analysis includes FC data from only 592 participants, this sample remains larger
394 than any clinical trial reported to date. Furthermore, the cohort was similar in terms of
395 reaction thresholds and extent of reproducibility to our previous report of over 3000
396 DBPCFC reported for peanut, an analysis which was robust using multiple sensitivity
397 analyses.² While three of the included studies utilized DBPCFC, a fourth (Blumchen et al¹⁴)
398 used open challenge methodology, otherwise conducted according to PRACTALL. Double-
399 blind methodology was not feasible in this study, since the challenge protocol used a
400 prolonged 2 hour dosing interval to specifically evaluate the duration of symptoms (which
401 would otherwise be obscured with shorter intervals). We did perform a sensitivity analysis
402 comparing the studies utilizing DBPCFC and the study by Blumchen et al, and did not
403 observe any differences (data not shown). Importantly, the inclusion of Blumchen et al also
404 allowed a more realistic evaluation of the any reduction in rate of anaphylaxis when using
405 less objective symptoms, since the prolonged challenge intervals reduced “carryover” of
406 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 only an average of one dosing interval results in a significant reduction in anaphylaxis at FC.

The assessment of reproducibility might be limited by the inclusion of three 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 Figure 3D, which is likely to reflect natural tolerance within some subjects allocated to the placebo arm in the included studies; however even in these studies the shift in threshold was no greater than that seen in the TRACE study performed in peanut-allergic adults: in the latter, 68% of participants had a shift in threshold limited to a $\frac{1}{2}$ -log change, compared to 67-74% of children in the 3 pediatric studies. While ideally this analysis would have included additional cohorts, we were unable to identify other cohorts where 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 this analysis simply omitted to sufficient detail regarding challenge-stop criteria in their publications.

Notwithstanding, our analysis is the first to compare the performance of the new proposed CoFAR version 3 challenge-stop criteria against PRACTALL, 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 significant (subjective) symptoms to be present from at least 2 organ systems. We present one 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.

ACKNOWLEDGEMENTS

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.

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 Dec;130(6):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 Jun;147(6):2249-2262.
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 Nov;150(5):1135-1143.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 Feb;32(2):342-348.
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 Feb;143(2):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 Nov 22;379(21):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 Oct;4(10):728-739.
8. Du Toit G, Roberts G, Sayre PH, Bahnson HT, Radulovic S, Santos AF, et al.; LEAP Study Team. Randomized trial of peanut consumption in infants at risk for peanut allergy. *N Engl J Med*. 2015 Feb 26;372(9):803-13.
9. Jones SM, Kim EH, Nadeau KC, Nowak-Wegrzyn A, Wood RA, Sampson HA, et al; Immune Tolerance Network. 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 Jan 22;399(10322):359-371.

- 475 10. Chinthrajah RS, Jones SM, Kim EH, Sicherer SH, Shreffler W, Lanser BJ, et al.
 476 Updating the CoFAR Grading Scale for Systemic Allergic Reactions in Food Allergy. *J*
 477 *Allergy Clin Immunol.* 2022 Jun;149(6):2166-2170.e1.
- 478 11. Fleischer DM, Greenhawt M, Sussman G, Begin P, Nowak-Wegrzyn A, Petroni D, et al.
 479 Effect of epicutaneous immunotherapy vs placebo on reaction to peanut protein
 480 ingestion among children with peanut allergy: the PEPITES randomized clinical trial.
 481 *JAMA* 2019;321:946-55.
- 482 12. Patel N, Vazquez-Ortiz M, Lindsley S, Campbell DE, Turner PJ. Low frequency of soya
 483 allergy in peanut-allergic children: relevance to allergen labelling on medicines. *Allergy*
 484 2018 Jun;73:1348-50.
- 485 13. Dua S, Ruiz-Garcia M, Bond S, Durham SR, Kimber I, Mills C, et al. Effect of sleep
 486 deprivation and exercise on reaction threshold in adults with peanut allergy: a
 487 randomized controlled study. *J Allergy Clin Immunol* 2019;144:1584-94.e2.
- 488 14. Blumchen K, Trendelenburg V, Ahrens F, Gruebl A, Hamelmann E, Hansen G, et al.
 489 Efficacy, Safety, and Quality of Life in a Multicenter, Randomized, Placebo-Controlled
 490 Trial of Low-Dose Peanut Oral Immunotherapy in Children with Peanut Allergy. *J Allergy*
 491 *Clin Immunol Pract.* 2019 Feb;7(2):479-491.
- 492 15. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).*
 493 2012;22(3):276-82.
- 494 16. Cardona V, Ansotegui IJ, Ebisawa M, El-Gamal Y, Fernandez Rivas M, Fineman S, et
 495 al. World allergy organization anaphylaxis guidance 2020. *World Allergy Organ J.* 2020
 496 Oct 30;13(10):100472.
- 497 17. Burks AW, Jones SM, Wood RA, Fleischer DM, Sicherer SH, Lindblad RW, et al;
 498 Consortium of Food Allergy Research (CoFAR). Oral immunotherapy for treatment of
 499 egg allergy in children. *N Engl J Med.* 2012 Jul 19;367(3):233-43.
- 500 18. Turner PJ, Patel N, Mäkelä MJ, Kukkonen K, Deschildre A, Blumchen K, et al.
 501 Improving the reporting of allergic adverse events during immunotherapy for food
 502 allergy. *J Allergy Clin Immunol.* 2022 Nov;150(5):1242-1244.
- 503 19. Frugier C, Graham F, Samaan K, Paradis L, Des Roches A, Bégin P. Potential Efficacy
 504 of High-Dose Inhaled Salbutamol for the Treatment of Abdominal Pain During Oral Food
 505 Challenge. *J Allergy Clin Immunol Pract.* 2021 Aug;9(8):3130-3137.
- 506 20. Dua S, Ruiz-Garcia M, Bond S, Dowey J, Durham SR, Kimber I, et al. Effects of
 507 Exercise and Sleep Deprivation on Reaction Severity During Oral Peanut Challenge: A

Randomized Controlled Trial. J Allergy Clin Immunol Pract. 2022 Sep;10(9):2404-2413.e1.

Journal Pre-proof

Table I: Characteristics of included cohorts.

Study	Published	N Data included:		Age of cohort	Inclusion criteria	DBPCFC protocol (mg peanut protein)	Challenge-stopping criteria	Median cumulative reaction dose at FC
		Baseline FC	Repeat FC					
PEPITES ¹¹	356	356	109	4-11y, median 7y	Reaction to ≤ 444 mg	DBPCFC, 30min intervals 1,3,10,30,100,300;	Based on PRACTALL	144mg (IQR 44-444)
BOPI Study ¹²	64	64	19	8-16y, median 13y	Reaction to ≤ 443 mg	DBPCFC, 30-60min intervals 3,10,30,100,300,1000,3000	PRACTALL	143mg (IQR 43-443)
TRACE Study ¹³	123	118	71	18-45y, mean 25y	Reaction to ≤ 1433 mg	DBPCFC, 30-60min intervals 0.003,0.03,0.3,3,30,100,300,1000	Modified PRACTALL	133mg (IQR 133-433)
Blumchen et al ¹⁴	62	54	24	3-17y median 7y	Reaction to ≤ 4500 mg	Open FC, 2h intervals Day 1: 3,10,30,100; Day 2: 100,300,1000,3000 Day 3: 4500	Objective symptoms according to PRACTALL	143mg (IQR 43-1400)

All doses expressed in mg peanut protein.

DBPCFC, double-blind, placebo-controlled food challenge; FC, food challenge; IQR, interquartile range.

518 **Table II:** Change in threshold at repeat FC by IPD meta-analysis, using different challenge-
 519 stop criteria (see text)

	Proportion of participants with	
	Maximum $\frac{1}{2}$ -log change in threshold	Maximum 1-log change in threshold
CoFAR mild x 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	71.2% (56.2-82.6%)	91.2% (84.1-95.3%)

520

521

FIGURE LEGENDS

Figure 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 DLS with PRACTALL and those utilising the framework used in the LEAP study. Any **red** symptom is generally considered to be dose-limiting. 3 **orange** symptoms are considered to imply a likely allergic reaction under PRACTALL, 2 **orange** symptoms under CoFAR v3 or the LEAP study DLS criteria. **Green** symptoms do not contribute towards the decision to terminate a food challenge.

*persisting symptom defined as ongoing beyond the duration of a dosing interval (minimum 20 minutes).

Figure 2: Proportion of peanut-allergic individuals in the combined cohort, tolerating a given dose, with DLS defined according to one of 3 definitions: (i) PRACTALL, (ii) at least 1 moderate symptom as per CoFAR v3, and (iii) a combination of 2+ mild-grade symptoms in ≥ 2 systems according to CoFAR v3. MTD, maximum tolerated dose.

Figure 3: Violin plot of the distributions of the log-change in reaction thresholds as a marker of the reproducibility of food challenges 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 the dotted lines represent the interquartile range. **A**, combined study cohort. Right panel, the change in threshold for each cohort, using the following challenge-stop definitions: **B**, a combination of 2 or more mild-grade symptoms in ≥ 2 systems according to CoFAR v3; **C**, at least 1 moderate symptom as per CoFAR v3; and **D**, PRACTALL.

Figure 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 three challenge-stop criteria assessed. Dashed lines represent the median, and the dotted lines represent the interquartile range.

Figure 5: Heatmap demonstrating the type of symptoms experienced at FC prior to (PRE-) and following (POST-) OIT in the 2 included studies. Transient symptoms defined as CoFAR grade 1 symptoms which resolve within 1 challenge dosing interval. CoFAR grade symptoms defined according to CoFAR grading system version 2.¹⁶

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

	PRACTALL	SCORPIO	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

