Prediction of pediatric peanut oral food challenge outcomes using machine learning

Jonathan Gryak, PhD,[a] Aleksandra Georgievska, BS,[a] Justin Zhang, BS,[b] Kayvan Najarian, PhD,[b,c,d,e,f] Rajan Ravikumar, MD,[g] Georgiana Sanders, MD MS,[g,h] and Charles F. Schuler IV, MD[g,h] *New York, NY, and Ann Arbor, Mich*

Background: Clinical testing, including food-specific skin and serum IgE level tests, provides limited accuracy to predict food allergy. Confirmatory oral food challenges (OFCs) are often required, but the associated risks, cost, and logistic difficulties comprise a barrier to proper diagnosis.
Objective: We sought to utilize advanced machine learning methodologies to integrate clinical variables associated with peanut allergy to create a predictive model for OFCs to improve predictive performance over that of purely statistical methods.
Methods: Machine learning was applied to the Learning Early about Peanut Allergy (LEAP) study of 463 peanut OFCs and associated clinical variables. Patient-wise cross-validation was used to create ensemble models that were evaluated on holdout test sets. These models were further evaluated by using 2 additional peanut allergy OFC cohorts: the IMPACT study cohort and a local University of Michigan cohort.
Results: In the LEAP data set, the ensemble models achieved a maximum mean area under the curve of 0.997, with a sensitivity and specificity of 0.994 and 1.00, respectively. In the combined validation data sets, the top ensemble model achieved a maximum area under the curve of 0.871, with a sensitivity and specificity of 0.763 and 0.980, respectively.
Conclusions: Machine learning models for predicting peanut OFC results have the potential to accurately predict OFC outcomes, potentially minimizing the need for OFCs while increasing confidence in food allergy diagnoses. (J Allergy Clin Immunol Global 2024;3:100252.)

Key words: *Peanut allergy, food allergy, anaphylaxis, oral food challenges, machine learning*

Abbreviations used
Ara h: *Arapis hypogaea*
AUC: Area under the curve
FA: Food allergy
LEAP: Learning Early about Peanut Allergy
LR: Logistic regression
LUCCK: Learning Using Concave and Convex Kernels
OFC: Oral food challenge
SHAP: Shapley Additive Explanations
SPT: Skin prick test
UM: University of Michigan

Food allergy (FA) affects 8% of children in the United States, and peanut allergy affects approximately 2.2%.[1] FA causes food anaphylaxis, leading to 200,000 US emergency room visits annually.[2,3] Diagnosis of FA currently relies heavily on patient history and oral food challenges (OFCs) because noninvasive FA diagnostics give false-positive rates of 30% to 50% in general and perhaps 90% among children with atopic dermatitis.[4-6] Unfortunately, patients and allergists may avoid OFCs because of the cost, perceived risk, or logistic constraints involved.[7] In addition, factors such as geography may affect OFC availability for a given individual.[8] Therefore, improvements to FA diagnosis that clarify the perceived risk and might expand the accessibility of allergy diagnoses are sorely needed.

Prior efforts have sought to combine FA-related variables to make predictive models for peanut OFCs. In DunnGalvin et al,[9] a logistic regression (LR) model was created by using 6 variables, namely, total and specific IgE levels, skin prick test (SPT) wheal size, symptom severity score, age, and sex. Although the LR model achieved good performance on a holdout set, the model relied primarily on the severity of prior symptoms, making it less useful in predicting OFCs without a clinical history. In a systematic review of various studies assessing the diagnostic accuracy of a variety of peanut allergy tests, such as SPT results, total specific IgE level, and IgE level in response to peanut components, Klemans et al[10] found quite varied but overall limited accuracy across a variety of testing contexts.

In prior work by members of our research group,[11] a machine learning methodology was developed on a retrospective data set from the University of Michigan (UM) containing features similar to those in prior work.[9] The work tested multiple machine learning models, such as Learning Using Concave and Convex Kernels (LUCCK)[12] and random forest, along with multiple feature selection and cross-validation strategies to create predictive models for peanut, egg, and milk challenges. The model for predicting peanut OFCs performed reasonably well, achieving an area under the curve (AUC), sensitivity, and specificity of 0.91, 0.89, and 0.92, respectively, on a holdout test set.

From [a]the Department of Computer Science, Queens College, City University of New York; [b]the Department of Computational Medicine and Bioinformatics, [c]the Department of Emergency Medicine, [d]the Department of Computer Science and Engineering, [e]the Michigan Institute for Data Science, [f]the Max Harry Weil Institute for Critical Care Research and Innovation, [g]the Division of Allergy and Immunology, Department of Internal Medicine, and [h]the Mary H. Weiser Food Allergy Center, University of Michigan, Ann Arbor.

Received for publication November 9, 2023; revised January 4, 2024; accepted for publication February 15, 2024.
Available online April 7, 2024.

Corresponding author: Jonathan Gryak, PhD, Department of Computer Science, Queens College, City University of New York, 65-30 Kissena Blvd, SB-A116, Queens, NY 11367. E-mail: jgryak@qc.cuny.edu.

2772-8293

https://doi.org/10.1016/j.jacig.2024.100252

Leveraging the framework developed previously, in this study we utilized 2 publicly available retrospective data sets from the UK-based Learning Early about Peanut Allergy (LEAP) study[13] and the IMPACT study,[14] as well as a separate, recent internal UM OFC data set,[15] all of which included variables such as serum-specific peanut component IgEs, to determine whether the addition of these now-common clinical tests would improve predictive performance. The resulting model achieved excellent test characteristics in the primary training/testing data set (the LEAP data set) and performed well in replication within the combined replication data sets (the IMPACT and UM data sets). Should the results of this use of this model be validated through the future application to prospective, multicenter data, the model has the potential to clarify the perceived risk to patients associated with OFCs while increasing confidence in FA diagnoses.

METHODS

Study cohorts

A secondary analysis of publicly available data via ITN Trial Share[16] that were collected as part of the LEAP study[13] was undertaken. The LEAP study was a single site randomized controlled trial to determine the effect of peanut avoidance or consumption on the development of peanut allergy in infants. Enrolled participants had to be between the ages of 4 and 11 months and be diagnosed with an egg allergy, severe eczema, or both conditions. Participants were randomly assigned to a peanut avoidance or consumption arm. Please note that each cohort measured the peanut skin test wheal size as the average of 2 perpendicular diameters per their respective protocols.[13,16]

An additional secondary analysis of similarly publicly available data[16] from the IMPACT study[14] was used in replication. The IMPACT study was a multicenter oral immunotherapy study in peanut allergy in an age group similar to that in the LEAP study. The entry OFCs were used in the present work.

Finally, all of the peanut OFCs from an internal multifood UM repository were deployed as a separate replication data set. This repository has been described previously.[15] Briefly, patients undergoing clinical OFCs performed in the UM Food Allergy Clinic were screened for enrollment, and baseline FA data and challenge outcome data were collected and deployed in this study.

OFCs

Each participant in the LEAP study was administered an OFC to assess peanut allergy at age 60 months. The challenges were a combination of open and double-blind, placebo-controlled challenges. Each participant in the IMPACT study underwent an entry double-blind, placebo-controlled food challenge before enrollment when within the age range of 12 to 48 months. Each participant in the internal UM data set underwent an open OFC. Before OFC administration, SPT and serum IgE level tests were performed and recorded in all studies.

Variable selection

Although the various data sets contain numerous data modalities, such as immunologic assays, genetic profiling, and metabolic panels, the variables chosen for inclusion in the machine learning model were those considered to be commonly collected and utilized in clinical practice. These include demographic features such as age, sex, and race; SPT results, including wheal and flare size measurements (flare if available); and serum IgE levels, including total IgE, peanut-specific IgE, and peanut component levels. In total, 15 variables were included, with the race categories other and unknown being dropped on account of low prevalence. Tables I to III provide the complete list of variables.

Machine learning

In Zhang et al,[11] a machine learning methodology was developed to predict OFCs for patients with suspected peanut, egg, or milk allergies. Multiple machine learning models and cross-validation strategies were evaluated, with the best-performing predictive models constructed by using an ensemble of models obtained via the LUCCK algorithm.[12] For this study, predictive models comprising ensembles of naive Bayes, support vector machine, random forest, and LUCCK models were trained and evaluated, one of which considered OFCs solely from patients in the avoidance arm of the LEAP study and the other of which utilized challenges from patients in both arms (ie, the avoidance and consumption arms). No separate model was made for the consumption-only arm owing to the low percentage of failed OFCs (3%) in that arm. Unlike in our previous model,[11] no additional feature selection beyond those in Table I was used. The data set was shuffled and split participant-wise into a 5-fold cross-validation data set (83%) and test data set (16%) and then stratified to preserve the pass-to-fail ratio of approximately 8:1, after which the predictive models were constructed. For the IMPACT and UM replication cohorts, the LEAP-trained machine learning models were applied directly. For a detailed description of the technical methods used, see the Supplemental Materials (available in the Online Repository at www.jaci-global.org).

Missing data

The UM study did not test for component *Arapis hypogaea* (Ara h) 9. The missing data were imputed by using the mean Ara h 9 values within the LEAP data set (0.170 kU/L). In 1 UM sample, all of the components (Ara h 1, Ara h 2, Ara h 3, Ara h 8, and Ara h 9) were missing and similarly imputed by using their respective LEAP study means. The IMPACT study did not collect flare, Ara h 8, or Ara h 9 data; as such, these values were imputed by using their associated LEAP study mean values of 2.43 mm, 1.23 kU/L, and 0.170 kU/L, respectively.

Model interpretation

Shapley Additive Explanations (SHAP) is a model interpretation method that uses the game-theoretic notion of Shapley values to determine the contribution of each feature to a model's output.[17] SHAP provides a means by which to elucidate a model's decision regarding individual samples and the aggregate importance of particular features. In the context of predicting OFCs, a positive Shapley value for a given feature and sample will contribute toward the model predicting an OFC pass, whereas negative values will contribute to the model predicting an OFC failure.

TABLE I. Demographic and clinical characteristics of patients in the LEAP group

Characteristic	Passed OFC (n = 412)	Failed OFC (n = 52)	*P* value
Age (mo), mean (95% CI)	7.78 (7.61-7.95)	7.61 (7.09 - 8.12)	.53
Sex (male), no. (%)	240 (58.3%)	36 (70.6%)	.09
Race, no. (%)			.46
White	306 (74.3%)	31 (60.8%)	
Black	31 (7.5%)	5 (9.8%)	
Asian	13 (3.2%)	2 (3.9%)	
Mixed	57 (13.8%)	10 (19.6%)	
Wheal size (mm), mean (95% CI)	0.41 (0.30-0.52)	10.04 (8.65-11.43)	<.001
Flare size (mm), mean (95% CI)	0.67 (0.47-0.87)	17.73 (15.60-19.85)	<.001
Total IgE level (kU/L), mean (95% CI)	499.84 (370.63-629.06)	868.51 (494.25-1242.77)	.08
Peanut IgE level (kU/L), mean (95% CI)	1.04 (0.62-1.46)	43.82 (24.04-63.59)	<.001
Ara h 1 level (kU/L), mean (95% CI)	0.07 (0.02-0.12)	17.38 (4.60-30.15)	.01
Ara h 2 level (kU/L), mean (95% CI)	0.04 (0.02-0.06)	36.88 (15.76-58.00)	.00
Ara h 3 level (kU/L), mean (95% CI)	0.14 (0.03-0.25)	4.58 (0.33-8.82)	.05
Ara h 8 level (kU/L), mean (95% CI)	1.10 (0.54-1.65)	5.47 (0.14-10.81)	.12
Ara h 9 level (kU/L), mean (95% CI)	0.28 (0.00-0.57)	0.73 (0.00-1.54)	.31

TABLE II. Demographic and clinical characteristics of patients in the IMPACT cohort

Characteristic	Passed OFC (n = 0)	Failed OFC (n = 140)	*P* value
Age (mo), mean (95% CI)	N/A	36.81 (35.30-38.33)	N/A
Sex (male), no. (%)	N/A	94 (67.14%)	N/A
Race, no. (%)			N/A
White	N/A	93 (66.43%)	
Black	N/A	6 (4.29%)	
Asian	N/A	15 (10.71%)	
Mixed	N/A	26 (18.57%)	
Wheal size (mm), mean (95% CI)	N/A	15.91 (14.97-16.86)	N/A
Flare size (mm), mean (95% CI)	N/A	N/A	N/A
Total IgE level (kU/L), mean (95% CI)	N/A	606.79 (471.85-741.73)	N/A
Peanut IgE level (kU/L), mean (95% CI)	N/A	133.93 (100.58-167.29)	N/A
Ara h 1 level (kU/L), mean (95% CI)	N/A	17.79 (13.23-22.35)	N/A
Ara h 2 level (kU/L), mean (95% CI)	N/A	70.29 (56.69-83.90)	N/A
Ara h 3 level (kU/L), mean (95% CI)	N/A	4.10 (2.75-5.45)	N/A
Ara h 8 level (kU/L), mean (95% CI)	N/A	1.23 (1.23-1.23)	N/A
Ara h 9 level (kU/L), mean (95% CI)	N/A	N/A	N/A

N/A, Not applicable.
No *P* values are available for this group because all of the participants reacted on food challenge.

Statistical comparison

LR models for OFC outcome prediction were constructed by using the same variables as used in the aforementioned machine learning methodology for each data set (Table I). Additionally, ensembles of LR models were created for a more commensurate comparison with the LUCCK ensemble models.

Study approval

The LEAP and IMPACT deidentified data sets are publicly available as already stated. The internal UM study was approved by the UM institutional review board under identifier HUM00165471, and all participants or their parent/guardian(s) provided informed consent; pediatric patients provided age-appropriate assent (assent was waived for those aged 6 or younger).

Data availability

The LEAP and IMPACT deidentified data sets are publicly available, as stated earlier. For the internal UM data set, all data pertain to human participants, and data requests would require a data transfer agreement subject to standard UM Data Office oversight, with the initial request directed to the corresponding author.

RESULTS

The LEAP data set used for this study's analysis contained 463 challenges, of which 52 (11.2%) resulted in reactions (ie, the patients undergoing the challenge experienced an adverse reaction), whereas all of the other challenged participants passed (ie, no adverse reaction was observed) (Table I). The mean age of initial enrollment was approximately 8 months, and all challenges were performed in patients aged 60 months. Compared with the nonreactors, those participants who experienced reactions had larger skin test wheals (10 mm vs 0 mm) and flares (18 mm vs 1 mm) as well as higher levels of peanut-specific IgE (43.8 vs 1.0 kU/L) and Ara h 2 IgE (36.9 vs 0.04 kU/L).

The IMPACT data set contained 140 entry OFCs, with all 140 of the participants who underwent OFC (100%) experiencing

TABLE III. Demographic and clinical characteristics of patients in the UM cohort

Characteristic	Passed OFC (n = 38)	Failed OFC (n = 8)	*P* value
Age (mo), mean (95% CI)	136.01 (92.07-179.95)	86.24 (44.42-128.05)	.14
Sex (male), no. (%)	19 (50.00%)	6 (75.00%)	.20
Race, no. (%)			.51
White	28 (73.68%)	7 (87.50%)	
Black	3 (7.89%)	0 (0.00%)	
Asian	4 (10.53%)	2 (25.00%)	
Mixed	1 (2.63%)	1 (12.50%)	
Wheal size (mm), mean (95% CI)	3.84 (2.77-4.92)	11.25 (7.54-14.96)	.01
Flare size (mm), mean (95% CI)	12.68 (8.12-17.25)	29.25 (19.11-39.39)	.02
Total IgE level (kU/L), mean (95% CI)	275.08 (54.00-496.15)	184.25 (54.75-313.75)	.50
Peanut IgE level (kU/L), mean (95% CI)	0.91 (0.49-1.33)	9.25 (0.00-21.23)	.24
Ara h 1 level (kU/L), mean (95% CI)	0.54 (0.05-1.02)	0.30 (0.00-0.70)	.48
Ara h 2 level (kU/L), mean (95% CI)	0.27 (0.11-0.42)	4.52 (0.00-11.32)	.29
Ara h 3 level (kU/L), mean (95% CI)	0.14 (0.09-0.19)	0.28 (0.00-0.60)	.44
Ara h 8 level (kU/L), mean (95% CI)	0.27 (0.09-0.45)	1.24 (0.00-3.05)	.36
Ara h 9 level (kU/L), mean (95% CI)	N/A	N/A	N/A

TABLE IV. Demographic and clinical characteristics of patients in the combined UM and IMPACT cohort

Characteristic	Passed OFC (n =38)	Failed OFC (n = 148)	*P* value
Age (mo), mean (95% CI)	136.01 (92.07-179.95)	39.49 (36.26-42.71)	<.001
Sex (male), no. (%)	19 (50.00%)	100 (67.57%)	.04
Race, no. (%)			.11
White	28 (73.68%)	100 (67.57%)	
Black	3 (7.89%)	6 (4.05%)	
Asian	4 (10.53%)	17 (11.49%)	
Mixed	1 (2.63%)	27 (18.24%)	
Wheal size (mm), mean (95% CI)	3.84 (2.77-4.92)	15.66 (14.73-16.59)	<.001
Flare size (mm), mean (95% CI)	12.68 (8.12-17.25)	3.88 (2.76-5.00)	<.001
Total IgE level (kU/L), mean (95% CI)	275.08 (54.00-496.15)	583.95 (455.19-712.71)	.02
Peanut IgE level (kU/L), mean (95% CI)	0.91 (0.49-1.33)	127.19 (95.31-159.07)	<.001
Ara h 1 level (kU/L), mean (95% CI)	0.54 (0.05-1.02)	16.84 (12.49-21.20)	<.001
Ara h 2 level (kU/L), mean (95% CI)	0.27 (0.11-0.42)	66.74 (53.64-79.83)	<.001
Ara h 3 level (kU/L), mean (95% CI)	0.14 (0.09-0.19)	3.89 (2.61-5.18)	<.001
Ara h 8 level (kU/L), mean (95% CI)	0.27 (0.09-0.45)	1.23 (1.13-1.32)	<.001
Ara h 9 level (kU/L), mean (95% CI)	N/A	N/A	N/A

Flare (for IMPACT cohort members) and Ara h 9 (for all cohort members) values are imputed from the LEAP study data, as earlier.

reactions (Table II). The mean age of enrollment was approximately 37 months, with all challenges occurring at enrollment. Because all of the participants had reactions, comparisons between reacting and nonreacting groups are not feasible; hence, the lack of *P* values in Table II. The mean skin test wheal size was 16 mm; flare was not reported in this study. The mean serum-specific peanut IgE level was 133.9 kU/L, and the mean Ara h 2 level was 70.3 kU/L.

The UM internal data set contained 46 patients who underwent OFC, of whom 8 (17%) experienced reactions (Table III). The mean age at challenge was approximately 127 months. Those with reactions had larger skin test wheals (11 mm vs 4 mm) and flares (29 mm vs 13 mm) than the nonreactors did, as well as higher levels of peanut-specific IgE (9.3 vs 0.9 kU/L) and Ara h 2 IgE (4.5 vs 0.04 kU/L).

Given the high rate of reactions in the IMPACT data set and the high rate of nonreactions in the UM data set, we combined these data sets to evaluate how the machine learning approaches would perform in a larger, aggregate data set containing a mix of higher-risk and lower-risk peanut OFCs covering various outcomes. This aggregate "combined" validation data set is summarized in Table IV; it contains data on a total of 186 OFCs, with a total of 148 reactions (80%).

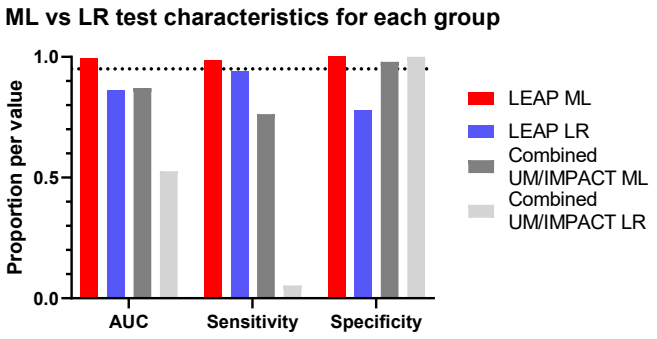

FIG 1. Summary of the performance metrics of the OFC outcome prediction models for the LEAP and combined UM-IMPACT data sets. The top machine learning (ML) approach versus the ensemble LR is shown for the 2 cohorts.

The results of the top-performing machine learning model versus the LR results in the LEAP and combined UM-IMPACT

TABLE V. Summary statistics of most accurate machine learning analysis outcomes by cohort

Data set/metric	AUC	F1	Accuracy	Sensitivity	Specificity	PPV	NPV	Best model
LEAP test data set (avoidance)	0.985	0.985	0.976	0.971	1	1	0.875	LUCCK ensemble
LEAP test data set (all)	0.993	0.993	0.987	0.986	1	1	0.9	LUCCK ensemble
UM validation	0.816	0.774	0.696	0.632	1	1	0.364	LUCCK ensemble
IMPACT validation (baseline)	N/A	0	0.993	N/A	0.993	0	1	SVM ensemble
UM-IMPACT combined data set	0.871	0.829	0.935	0.763	0.980	0.906	0.942	Naive Bayes ensemble
% Difference between the LEAP and UM-IMPACT data sets	–12%	–17%	–5%	–23%	–2%	–9%	5%	

NPV, Negative predictive value; *PPV,* positive predictive value; *SVM,* support vector machine. The bolded datasets were used in the calculation of the % difference.

data sets are each presented in Fig 1 and in Table V. The LUCCK ensemble model performed best in the LEAP data set, providing an AUC of 0.993, sensitivity of 0.986, and specificity of 1.000. These results compared favorably with those provided by the LR ensemble in LEAP, which gave an AUC of 0.860, sensitivity of 0.942, and specificity of 0.778. In the replication combined UM-IMPACT data set, the naive Bayes ensemble machine learning model performed best, providing an AUC of 0.871, sensitivity of 0.763, and specificity of 0.980. These results also compared favorably with those provided by the relevant LR ensemble, which provided an AUC of 0.526, sensitivity of 0.053, and specificity of 1.000. Full details regarding each machine learning model's performance in each group are available in Table E1 (see the Online Repository at www.jaci-global.org).

The predictive model for the LEAP and combined UM-IMPACT cohorts were each analyzed via SHAP to understand the data components on which the model relied to make decisions. Fig 2 presents the mean absolute Shapley values for each cohort, which allow us to investigate the relative contribution of each feature irrespective of class. On average in LEAP (Fig 2, *A*), flare size had the highest mean Shapley value, closely followed by wheal size. Rounding out the top 5 in decreasing order of mean values were total IgE level, serum-specific peanut IgE level, and age. Component peanut IgE features such as Ara h 3 and Ara h 9 comprised the lowest-contributing factors, with the exception of Asian race, which ranked 14th among the 15 features in terms of overall contribution. In the combined UM-IMPACT cohort (Fig 2, *B*), wheal size and Ara h 2 IgE were the top 2 contributors, and flare ranked third. Interestingly, total IgE level was much less important than peanut-specific IgE and/or component level testing and skin test results.

The per-sample SHAP analyses for the top 4 features—flare size, wheal size, total IgE level, and peanut IgE levels—are presented in Fig 3 for both the LEAP and combined UM-IMPACT cohorts. In the graphs in Fig 3, each point represents a sample in the test set, with the blue points representing those samples that the model predicts will not react in OFC, and the red points being those that the model predicts will result in a reactive OFC. The x-axis corresponds to Shapley value, with positive values indicating that the given feature contributes toward the model making a prediction of OFC pass and negative values associated with predictions of OFC failure. The y-axis corresponds to the numeric value of the feature. For example, in Fig 3, which compares model predictions with wheal size, one can observe that large wheal sizes have negative Shapley values, which indicates that on the basis of wheal size alone, the model would predict a failed OFC, and this is indeed the prediction made by the model for many samples, as indicated by the red points.

Fig 4 displays the degree to which each data point correlated positively or negatively with all of the other data points included in the machine learning model. Of note, the SPT wheal and flare size results were positively correlated with each other as well as with peanut-specific total IgE and Ara h 1, Ara h 2, and Ara h 3 IgE level results. Consistent with the SHAP analysis, both SPT wheal and flare size correlated strongly with a reaction on OFC.

DISCUSSION

Given the known limitations of current testing paradigms, interest in new diagnostic approaches for FA is strong.[18] Although attempts have been made to use advanced statistical methods to enhance the accuracy of prediction of peanut OFC outcome, these attempts have rarely result in AUC values greater than 0.90.[10,19] In the present study, even a well-defined clinical trial population (the LEAP population) produced an AUC of only 0.85 with use of LR ensemble models for OFC outcome (see Table E1). The relatively poor performance of SPT and IgE level results alone in predicting OFC outcome, particularly in the context of equivocal histories of allergy, has prompted attempts to develop novel test modalities.[20,21] Novel FA diagnostic approaches include component-resolved diagnostics,[22,23] *ex vivo* allergen activation of allergic cell types,[24-28] and epitope mapping.[29] Although these are of great interest and potential, various logistic or biologic aspects of these tests make their widespread adoption difficult.

Machine learning–based approaches provide 2 potential advantages to add to FA diagnostic development. First, machine learning can rely on existing data sets, meaning that secondary use of extant data can be reanalyzed by using machine learning, as has been done herein. Second, machine learning can be applied to new data sets as they are developed, meaning that nearly any approach to FA can be folded into a machine learning context to test the additive effect of new test modalities. Furthermore, if machine learning approaches can become successful, the increasing ubiquity of internet access could result in relatively rapid adoption at the clinical level, avoiding the logistic and biologic hurdles mentioned for advanced testing via other means.

To our knowledge, this study is the first in FA to use machine learning on a well-reported, well-controlled, clinical trial data set to quantify OFC likelihood based on readily available clinical parameters. This study can be viewed in part as validating the prior work from our group in Zhang et al[11] and also as demonstrating how machine learning models trained on one FA data set may be made readily applicable to other data sets. We are aware of a similar report of machine learning used in cooked egg allergy[30]; that study is notable as an analysis of a clinical retrospective cooked egg challenge cohort, although

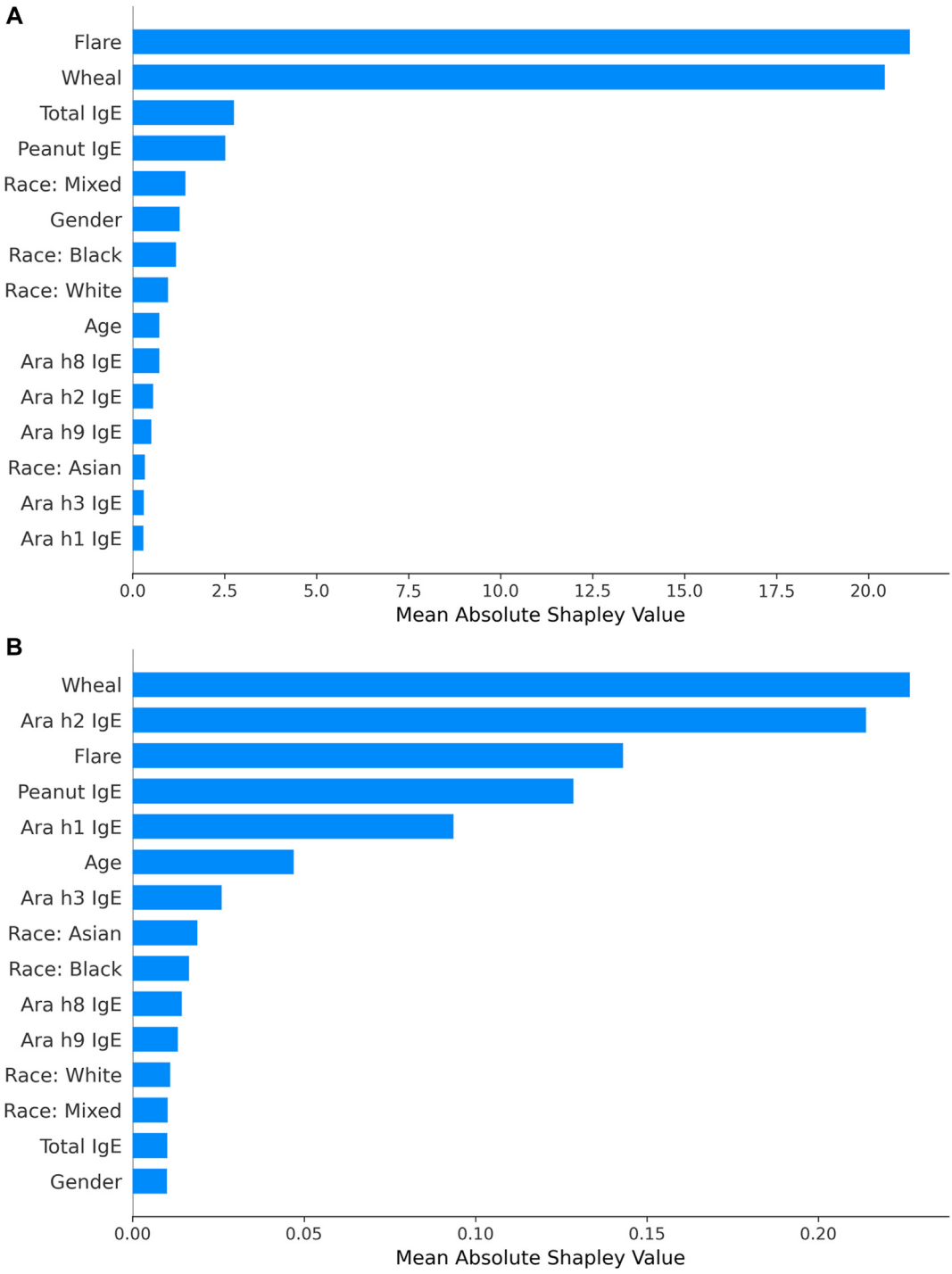

FIG 2. Mean absolute Shapley values for the final ensemble machine learning model with the greatest accuracy for the LEAP cohort (**A**) and combined UM-IMPACT cohort (**B**). Higher values for a given feature indicate a greater contribution toward the model's prediction over the test set.

the predictive power of the machine learning model was relatively low, perhaps partly because of the total number of OFCs analyzed (n = 67). Machine learning has been applied in other FA contexts. One intriguing work described analyzing the infant gut microbiome by using machine learning–based approaches to define ultimate FA risk.[31] Another study focused on epigenetic markers associated with FA, although this work was not clearly defined by OFC.[32] A final study used machine learning to evaluate physician diagnoses of FA in the Canadian Primary Care Sentinel Surveillance Network, but that study was not necessarily defined by OFC either.[33] Given the limited reports on machine learning in FA, multiple experts have called for machine

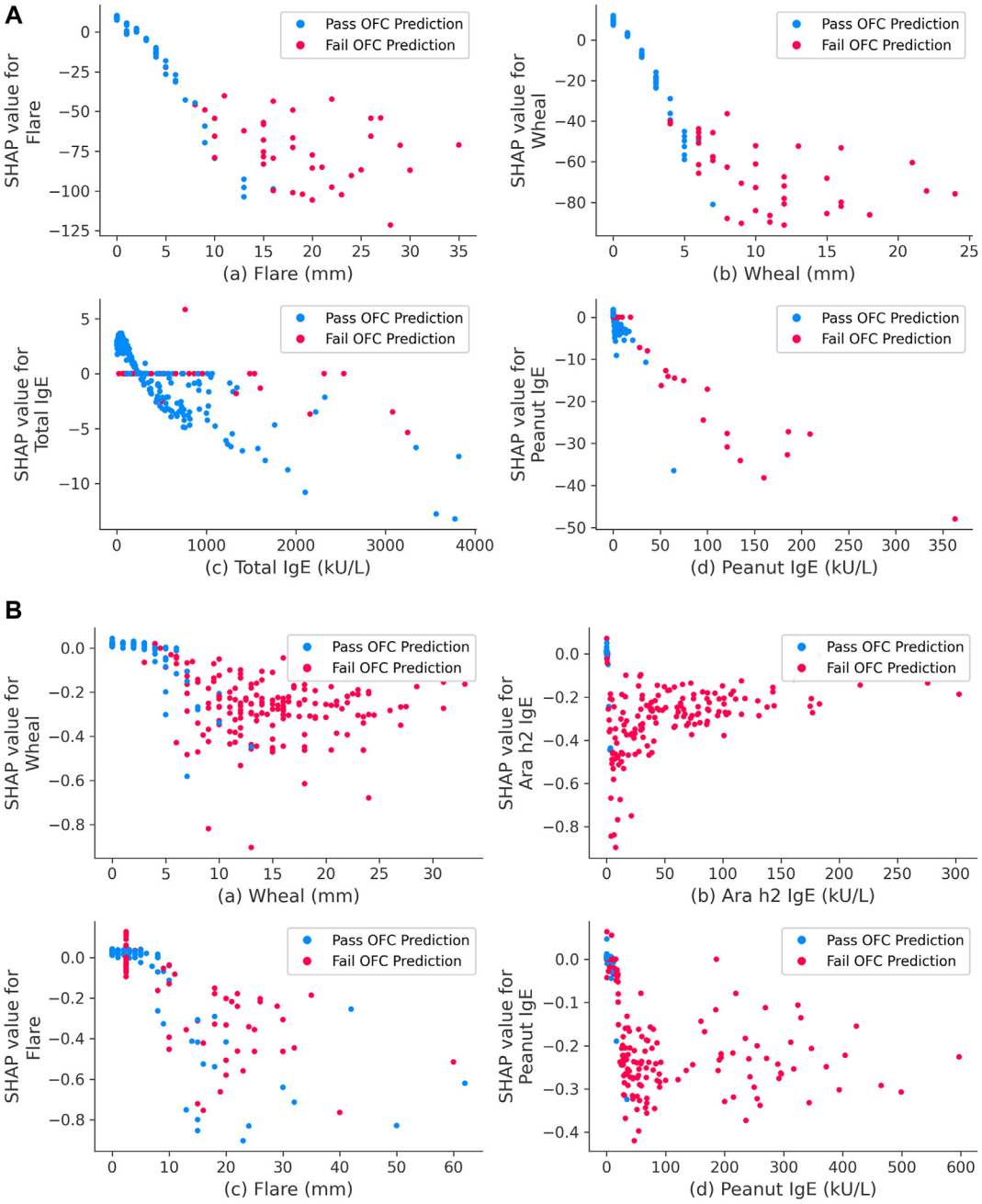

FIG 3. A comparison of Shapley values for the top 4 features each for the LEAP cohort (**A**) and combined UM-IMPACT cohort (**B**). Note that these are shown in order of model reliance on the feature in question, such that there is a different order in (**A**) versus in (**B**).

learning–based approaches to be applied in FA and to multi-omics approaches related to FA.[34-36] The current work helps to answer this call.

One key goal of this study was to determine whether a machine learning model trained on a specific data set can accurately predict outcomes in another data set. The LEAP-trained and LEAP-tested machine learning model provided outstanding accuracy in predicting LEAP OFC outcomes. Crucially, machine learning approaches trained on the LEAP data set still provided good predictive capacity when applied to the combined UM-IMPACT data set. Given that these 3 data sets are fundamentally different, the generalizability between the data sets is a very useful observation from this work. For example, the fact that open and double-blinded OFCs results were used in the different studies shows that these models may be able to generalize between result types. These results suggest that a larger, multisite, coordinated study has potential to develop a trained machine learning model that might provide a clinically usable prediction approach for OFC outcomes.

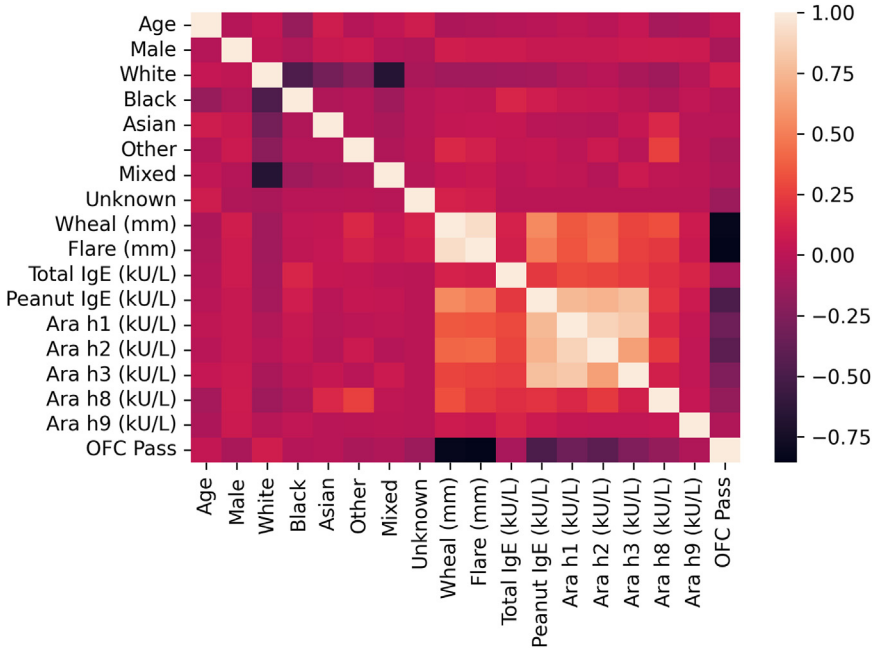

FIG 4. Correlation heatmap of all values included in the LUCCK machine learning model for LEAP.

Although skin testing alone may not fully predict OFC outcomes, in the machine learning context such data points can contribute to accurate predictions. Flare size was the most contributive feature in predicting OFC outcome as determined by SHAP for the LEAP cohort. This corroborates the finding in Zhang et al,[11] which utilized a data set distinct from that used in this study. As shown in Fig 3, *A*, larger flare sizes are associated with adverse OFC outcomes, with all cases having a flare size of at least 17 mm resulting in reactive OFCs. The relatively high importance of flare size is intriguing, given the low emphasis placed on flare sizes in predicting peanut OFC results in clinical practice.[4,37] In addition, the importance of wheal size as determined by SHAP can be readily understood by looking at the individual values presented in Fig 3. Reactive OFCs are associated with higher wheal sizes, with all of the OFCs for which the wheal was at least 8 mm in size corresponding to OFC failures; this reproduces the cutoff reported in Roberts et al.[38] For wheal sizes of 5 mm and lower, there is a plurality of OFCs (clustered between Shapley values of 0.0 and 0.4), for which wheal size contributes to a prediction of a passed OFC; yet, there are still a sizeable number of OFCs (the cluster between Shapley values of –20 and –5) for which the wheal size suggests an OFC failure. When applied to the combined UM-IMPACT data set, similar, but not identical, patterns emerge. Although flare size was not the top feature in this data set, it did rank third, still showing some importance. Whether the potential relative importance of flare size bears out in future data sets remains to be seen, but whether flare size does hold predictive meaning, defining its role could help improve OFC prediction in the future.

One potential concern might be whether the highly correlated values of wheal size and flare size each improve the model independently. Individual models trained via LUCCK mitigate this issue by determining the importance of similarity in each feature individually and by class, allowing for a better integration of features derived from similar and different modalities. Moreover, the use of ensemble models, which were the highest-performing models on both the LEAP and combined UM-IMPACT data sets, greatly reduces (but does not entirely eliminate) the likelihood of overfitting,[39] especially when trained against fewer training examples, which is a major issue in other complex machine learning methods for OFC prediction.

This study has several limitations. The data utilized in this retrospective study were for a single food. In addition, the validation cohort was amalgamated from 2 data sets to better cover a range of OFC outcomes. As such, validation utilizing a multicenter data set with standardized data inputs and minimal missing data, as well as in multiple food allergies, is necessary to ensure that the models can be generalizable. Furthermore, given that various factors such as skin tone, "sensitive skin" status, and UV light exposure may affect the measurement of wheal and flare size, future work will require attention to how consistency of SPT results might affect these models' outputs.[40] In addition, the peanut introduction or avoidance methods (either condition was pursued until age 60 months) in the LEAP study could produce results that differ from real-world food introductions. Confirmation of these results by using a separate data set with more typical food introduction methods will address this potential confounder. Moreover, the LEAP study inclusion criteria required participants to have either an egg allergy, severe eczema, or both, which likely led to the observed higher prevalence of peanut allergy in the study population (11%) than in the general population (2%).[41] Additional data modalities, such as prior ingestions or reactions, the presence of comorbid allergic disease (including atopic dermatitis and/or multiple food allergies), family history, patient-reported symptoms, novel diagnostics, and genomic information, were not included in the models. Such modalities could improve model performance and should be included in future models for OFC prediction.

In the future, comparing the outcome of a refined machine learning approach to predictions made by trained, experienced

allergists and then conducting OFCs prospectively would be the most definitive method to decide whether a machine learning–based approach can accurately predict OFC outcomes and potentially identify which features should be included and/or weighted to predict those outcomes. In addition, future work to determine the most accurate machine learning algorithm and approach would require a software system implementation. This would likely include the creation of a website and/or smartphone app to process inputs and deliver validated predictions, followed by testing in the appropriate context. This might include a primary care–focused or underserved region(s) study to define whether the approach could actually deliver an expansion of accurate allergy care beyond the allergy office.

In conclusion, although machine learning–based approaches to FA diagnosis hold promise, limited reports of such approaches exist. The present work demonstrates that machine learning–based approaches using clinically available FA data can provide accurate assessments of OFC risk and may improve on the accuracy gained via standard statistical models. Novel diagnostics, genomic and epigenomic data, and perhaps as-yet undefined factors may further enhance the potential accuracy of this approach. Future work to define machine learning as a diagnostic contributor in populations with allergy to other foods and more heterogeneous FA populations is needed, as the models, features, and/or weighting of features may ultimately vary by food allergen or other features. In addition, future work may seek to deploy machine learning in predicting the severity of an OFC reaction and not just the likelihood of a reaction. The approach used here can serve as a blueprint for the necessary work that can follow.

DISCLOSURE STATEMENT

Supported by a Michigan Food Allergy Research Accelerator grant from the University of Michigan's Mary H. Weiser Food Allergy Center, the Gerber Foundation (to C.S.), and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (grant K23AI162661 [to C.S.]).

Disclosure of potential conflict of interest: The authors declare that they have no relevant conflicts of interest.

We wish to acknowledge the LEAP study team, IMPACT study team, and Immune Tolerance Network, whose work and data-sharing practices made this work possible.

REFERENCES

1. Gupta RS, Warren CM, Smith BM, Blumenstock JA, Jiang J, Davis MM, et al. The public health impact of parent-reported childhood food allergies in the United States. Pediatrics 2018;142:e20181235.
2. Robinson LB, Arroyo AC, Faridi MK, Rudders S, Camargo CA. Trends in US emergency department visits for anaphylaxis among infants and toddlers: 2006-2015. J Allergy Clin Immunol Pract 2021;9:1931-8.
3. Rudders SA, Banerji A, Vassallo MF, Clark S, Camargo CA. Trends in pediatric emergency department visits for food-induced anaphylaxis. J Allergy Clin Immunol 2010;126:385-8.
4. Koplin JJ, Perrett KP, Sampson HA. Diagnosing peanut allergy with fewer oral food challenges. J Allergy Clin Immunol Pract 2019;7:375-80.
5. Peters RL, Allen KJ, Dharmage SC, Tang ML, Koplin JJ, Ponsonby AL, et al. Skin prick test responses and allergen-specific IgE levels as predictors of peanut, egg, and sesame allergy in infants. J Allergy Clin Immunol 2013;132:874-80.
6. Fleischer DM, Bock SA, Spears GC, Wilson CG, Miyazawa NK, Gleason MC, et al. Oral food challenges in children with a diagnosis of food allergy. J Pediatr 2011;158:578-83.
7. Greiwe J, Oppenheimer J, Bird JA, Fleischer DM, Pongracic JA, Greenhawt M, et al. AAAAI Work Group Report: trends in oral food challenge practices among allergists in the United States. J Allergy Clin Immunol Pract 2020;8:3348-55.
8. El Baba A, Jeimy S, Soller L, Kim H, Begin P, Chan ES. Geographical discrepancy in oral food challenge utilization based on Canadian billing data. Allergy Asthma Clin Immunol 2023;19:5.
9. DunnGalvin A, Daly D, Cullinane C, Stenke E, Keeton D, Erlewyn-Lajeunesse M, et al. Highly accurate prediction of food challenge outcome using routinely available clinical data. J Allergy Clin Immunol 2011;127:633-9.
10. Klemans RJ, van Os-Medendorp H, Blankestijn M, Bruijnzeel-Koomen CA, Knol EF, Knulst AC. Diagnostic accuracy of specific IgE to components in diagnosing peanut allergy: a systematic review. Clin Exp Allergy 2015;45:720-30.
11. Zhang J, Lee D, Jungles K, Shaltis D, Najarian K, Ravikumar R, et al. Prediction of oral food challenge outcomes via ensemble learning. Informatics in Medicine Unlocked 2022;101142.
12. Sabeti E, Gryak J, Derksen H, Biwer C, Ansari S, Isenstein H, et al. Learning using concave and convex kernels: applications in predicting quality of sleep and level of fatigue in fibromyalgia. Entropy 2019;21:442.
13. 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.
14. 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.
15. Schuler CFt, O'Shea KM, Troost JP, Kaul B, Launius CM, Cannon J, et al. Transepidermal water loss rises before food anaphylaxis and predicts food challenge outcomes. J Clin Invest 2023;133:e168965.
16. Immune Tolerance Network TrialShare home page. Availale at: https://www.itntrialshare.org. Accessed May 23, 2023.
17. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 30, 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), December 4-9, 2017, Long Beach, California, USA.
18. Santos AF, Kulis MD, Sampson HA. Bringing the next generation of food allergy diagnostics into the clinic. J Allergy Clin Immunol Pract 2022;10:1-9.
19. Sindher S, Long AJ, Purington N, Chollet M, Slatkin S, Andorf S, et al. Analysis of a large standardized food challenge data set to determine predictors of positive outcome across multiple allergens. Front Immunol 2018;9:2689.
20. Santos AF, Douiri A, Cares N, Wu SY, Stephens A, Radulovic S, et al. Basophil activation test discriminates between allergy and tolerance in peanut-sensitized children. J Allergy Clin Immunol 2014;134:645-52.
21. Hemmings O, Niazi U, Kwok M, James LK, Lack G, Santos AF. Peanut diversity and specific activity are the dominant IgE characteristics for effector cell activation in children. J Allergy Clin Immunol 2021;148:495-505.
22. Labrosse R, Graham F, Caubet JC. Recent advances in the diagnosis and management of tree nut and seed allergy. Curr Opin Allergy Clin Immunol 2022;22:194-201.
23. Abrams EM, Chan ES, Sicherer S. Peanut allergy: new advances and ongoing controversies. Pediatrics 2020;145:e20192012.
24. Santos AF, Couto-Francisco N, Bécares N, Kwok M, Bahnson HT, Lack G. A novel human mast cell activation test for peanut allergy. J Allergy Clin Immunol 2018;142:689-91.
25. Hemmings O, Du Toit G, Radulovic S, Lack G, Santos AF. Ara h 2 is the dominant peanut allergen despite similarities with Ara h 6. J Allergy Clin Immunol 2020;146:621-30.
26. Bahri R, Custovic A, Korosec P, Tsoumani M, Barron M, Wu J, et al. Mast cell activation test in the diagnosis of allergic disease and anaphylaxis. J Allergy Clin Immunol 2018;142:485-96.
27. Marrs T, Brough HA, Kwok M, Lack G, Santos AF. Basophil CD63 assay to peanut allergens accurately diagnoses peanut allergy in patient with negative skin prick test and very low specific IgE. Pediatr Allergy Immunol 2022;33:e13739.
28. Santos AF, Alpan O, Hoffmann HJ. Basophil activation test: Mechanisms and considerations for use in clinical trials and clinical practice. Allergy 2021;76:2420-32.
29. Suárez-Fariñas M, Suprun M, Kearney P, Getts R, Grishina G, Hayward C, et al. Accurate and reproducible diagnosis of peanut allergy using epitope mapping. Allergy 2021;76:3789-97.
30. Kuniyoshi Y, Tokutake H, Takahashi N, Kamura A, Yasuda S, Tashiro M. Machine learning approach and oral food challenge with heated egg. Pediatr Allergy Immunol 2021;32:776-8.
31. Metwally AA, Yu PS, Reiman D, Dai Y, Finn PW, Perkins DL. Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via long short-term memory networks. PLoS Comput Biol 2019;15:e1006693.
32. Alag A. Machine learning approach yields epigenetic biomarkers of food allergy: a novel 13-gene signature to diagnose clinical reactivity. PLoS One 2019;14:e0218253.
33. Singer AG, Kosowan L, Soller L, Chan ES, Nankissoor NN, Phung RR, et al. Prevalence of physician-reported food allergy in Canadian children. J Allergy Clin Immunol Pract 2021;9:193-9.

34. Czolk R, Klueber J, Sørensen M, Wilmes P, Codreanu-Morel F, Skov PS, et al. IgE-mediated peanut allergy: current and novel predictive biomarkers for clinical phenotypes using multi-omics approaches. Front Immunol 2020;11:594350.
35. Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. Leveraging multilayered "omics" data for atopic dermatitis: a road map to precision medicine. Front Immunol 2018;9:2727.
36. Proper SP, Azouz NP, Mersha TB. Achieving precision medicine in allergic disease: progress and challenges. Front Immunol 2021;12:720746.
37. Greenhawt M, Shaker M, Wang J, Oppenheimer JJ, Sicherer S, Keet C, et al. Peanut allergy diagnosis: a 2020 practice parameter update, systematic review, and GRADE analysis. J Allergy Clin Immunol 2020;146:1302-34.
38. Roberts G, Lack G. of Parents ALS, Team CS, and others. Diagnosing peanut allergy with skin prick and specific IgE testing. J Allergy Clin Immunol 2005;115:1291-6.
39. Bartlett P, Freund Y, Lee WS, Schapire RE. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat 1998;26:1651-86, 36.
40. Bernstein IL, Li JT, Bernstein DI, Hamilton R, Spector SL, Tan R, et al. Allergy diagnostic testing: an updated practice parameter. Ann of Allergy Asthma Immunol 2008;100:S1-148.
41. Akuete K, Guffey D, Israelsen RB, Broyles JM, Higgins LJ, Green TD, et al. Multicenter prevalence of anaphylaxis in clinic-based oral food challenges. Ann of Allergy Asthma Immunol 2017;119:339-48.