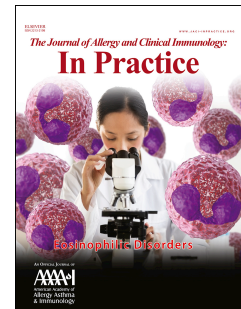


Journal Pre-proof

Combining allergen components improves the accuracy of peanut allergy diagnosis

Oliver Hemmings, MSc, Umar Niazi, MBBS PhD, Matthew Kwok, BSc, Suzana Radulovic, MD, George Du Toit, MBBS FRCPCH, Gideon Lack, MBBS FRCPCH, Alexandra F. Santos, MD PhD



PII: S2213-2198(21)00975-2

DOI: <https://doi.org/10.1016/j.jaip.2021.08.029>

Reference: JAIP 3817

To appear in: *The Journal of Allergy and Clinical Immunology: In Practice*

Received Date: 17 May 2021

Revised Date: 3 August 2021

Accepted Date: 25 August 2021

Please cite this article as: Hemmings O, Niazi U, Kwok M, Radulovic S, Du Toit G, Lack G, Santos AF, Combining allergen components improves the accuracy of peanut allergy diagnosis, *The Journal of Allergy and Clinical Immunology: In Practice* (2021), doi: <https://doi.org/10.1016/j.jaip.2021.08.029>.

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.

© 2021 Published by Elsevier Inc. on behalf of the American Academy of Allergy, Asthma & Immunology

Combining allergen components improves the accuracy of peanut allergy diagnosis

1 Oliver Hemmings, MSc^{1,2,3}; Umar Niazi, MBBS PhD⁴; Matthew Kwok, BSc^{1,2,3}; Suzana Radulovic,
2 MD^{1,2,5}; George Du Toit, MBBS FRCPCH^{1,2,5}; Gideon Lack, MBBS FRCPCH^{1,2,3,5}; Alexandra F.
3 Santos, MD PhD^{1,2,3,5*}

4 ¹Department of Women and Children's Health (Pediatric Allergy), School of Life Course
5 Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United
6 Kingdom

7 ²Peter Gorer Department of Immunobiology, School of Immunology and Microbial Sciences,
8 King's College London, London, United Kingdom

9 ³Asthma UK Centre in Allergic Mechanisms of Asthma, London, United Kingdom

10 ⁴Guy's and St Thomas' National Health Service Foundation Trust and King's College London
11 National Institute for Health Research Biomedical Research Centre Translational Bioinformatics
12 Platform, Guy's Hospital, London, UK.

13 ⁵Children's Allergy Service, Evelina London Children's Hospital, Guy's and St Thomas' Hospital,
14 London, United Kingdom

15 ***Corresponding Author:**

16 Alexandra F. Santos

17 Address: Department of Paediatric Allergy, 2nd floor, South Wing, St Thomas' Hospital, SE1 7EH
18 London, United Kingdom

19 Telephone number: +44 (0) 20 7188 6424

20 Fax number: +44 (0) 20 7403 8640

21 Email address: alexandra.santos@kcl.ac.uk

22

Funding: This work was supported by the Medical Research Council (MRC Clinician Scientist Fellowship MR/M008517/1, MRC Centenary Early Career Award and MRC Clinical Research Training Fellowship G0902018 awarded to A.F. Santos), Asthma UK (AUK-BC-2015-01) and by the UK National Institute for Health Research comprehensive Biomedical Research Centre award to Guy's and St Thomas' National Health Service (NHS) Foundation Trust, in partnership with the King's College London and King's College Hospital NHS Foundation Trust.

Conflicts of Interest (COI): Dr. Du Toit reports grants from National Institute of Allergy and Infectious Diseases (NIAID, NIH), grants from Food Allergy & Research Education (FARE), other from MRC & Asthma UK Centre, other from UK Dept of Health through NIHR, grants from Aimmune, grants from DBV Technologies, grants from National Peanut Board (USA), during the conduct of the study; other from FoodMaestro, outside the submitted work. Dr. Lack reports grants from National Institute of Allergy and Infectious Diseases (NIAID, NIH), other from Food Allergy & Research Education (FARE), other from MRC & Asthma UK Centre, other from UK Dept of Health through NIHR, other from National Peanut Board (NPB), other from The Davis Foundation, during the conduct of the study; personal fees and other from DBV Technologies, other from Mission Mighty Me, personal fees from Novartis, personal fees from Sanofi-Genzyme, personal fees from Regeneron, personal fees from ALK-Abello, personal fees and other from Lurie Children's Hospital, outside the submitted work; Dr. Santos reports grants and personal fees from Medical Research Council (G0902018, MR/M008517/1 and MR/T032081/1); grants from Asthma UK and the NIHR through the Biomedical Research Centre (BRC) award to Guy's and St Thomas' NHS Foundation Trust, during the conduct of the study; grants from Immune Tolerance Network/National Institute of Allergy and Infectious Diseases (NIAID, NIH), grants from Asthma UK; personal fees from Thermo Scientific, Nutricia, Infomed, Novartis, Allergy Therapeutics, Buhlmann, as well as research support from Buhlmann and Thermo Fisher Scientific through a collaboration agreement with King's College London. The other authors have nothing to disclose.

Word count: 3636

Tables: 2

Figures: 5

Online repository: 1 Table; E-methods with scripts

Abstract:

Background: IgE to peanut often occurs in the absence of peanut allergy. Detection of allergen component sIgE has improved diagnosis and birthed molecular allergen component arrays, in which sensitisation to multiple allergen components can be measured simultaneously.

Objective: To improve the diagnostic utility of serology for peanut allergy, by mapping interactions of sIgE to multiple components and IgE functional characteristics.

Methods: A cohort of 100 children was studied, with a 60-children cohort employed for external validation. Levels of total IgE, sIgE to peanut and peanut components were measured using singleplex ImmunoCAP and multiplex ISAC. Peanut IgE specific activity, avidity and diversity were determined. Diagnostic modelling was performed using a Bayesian hierarchical model.

Results: Sensitization to the 112 allergens on ISAC (model 1) demonstrated the highest accuracy to diagnose peanut allergy (AUC=0.92). Sensitization to peanut components on ISAC (model 2) reported an AUC of 0.86 and on singleplex (model 3) an AUC of 0.92, which was greater than that of Ara h 2-sIgE alone (AUC= 0.90). Functional characteristics of peanut-sIgE (model 4) reported an AUC of 0.89, which was in turn greater than that of peanut-sIgE (AUC=0.75). Model 3 offered the highest predictive value and the second highest overall diagnostic accuracy.

Conclusions: sIgE to a combination of allergen components (Ara h 1, 2, 3 and 6) is highly predictive of peanut allergy and superior to individual markers. Combining the functional characteristics of IgE was superior to peanut-sIgE levels alone. These models can be applied in real-time during clinical consultations using online calculators.

Abstract word count: 250

Highlights:

1. What is already known about this topic?

Specific IgE to peanut components, particularly Ara h 2, can be more accurate than specific IgE to peanut extract.

2. What does this article add to our knowledge?

Combining specific IgE to various peanut allergens or combining functional characteristics of IgE can improve the diagnostic accuracy of individual allergen-specific IgE or peanut-specific IgE levels, respectively.

3. How does this study impact current management guidelines?

Online calculators will be available to clinicians to calculate the probability of peanut allergy based on a combination of IgE to peanut allergens or of functional characteristics of peanut-specific IgE that can aid clinical decision-making.

Key words:

Singleplex, Multiplex, coefficient, specificity, specific activity, avidity, diversity, specific IgE, peanut allergy

Abbreviations:

AUC - Area Under the Curve

(ImmunoCAP) ISAC - Immuno Solid-phase Allergen Chip

NA – Non-sensitized Non-allergic

CV – Cross-validation

OFC - Oral Food Challenge

PA - Peanut Allergic

PS - Peanut Sensitized but Tolerant

ROC - Receiver Operating Characteristic

sIgE - Specific IgE

SPT - Skin Prick Test

109 Introduction:

110 Diagnosing peanut allergy is still a clinical conundrum in some cases, given the frequent
 111 discrepancy between detection of peanut-specific IgE and clinical peanut allergy. This
 112 discrepancy has been documented in various studies around the world, from birth cohort
 113 studies, outpatient clinic observational studies and interventional studies¹⁻³. The diagnosis of
 114 peanut allergy has improved with the detection of allergen component sIgE, such as Ara h 2-
 115 sIgE and Ara h 6-sIgE⁴⁻⁶. We recently showed that IgE to these components have 82% sensitivity
 116 and 94% and 90% specificity, respectively⁵. This emergence of component-resolved diagnostics
 117 (CRD) gave prominence to molecular allergen component arrays, such as the ImmunoCAP ISAC,
 118 in which sensitisation to multiple allergen components can be measured simultaneously, in a
 119 semi-quantitative fashion. This method has proved comparable to singleplex ImmunoCAP,
 120 albeit data is reported in arbitrary units, with some studies suggesting lower sensitivity^{7, 8}.
 121 Notably, using ISAC arrays adds important information on both cross and co-sensitisation to
 122 different allergens from multiple allergen groups.

123 Despite the improvements in the diagnosis of peanut allergy with CRD, there is still a
 124 substantial proportion of children being assessed for possible peanut allergy who require an
 125 oral food challenge (OFC) to clarify the allergic status^{3, 9}. We have developed and validated
 126 cellular tests using flow cytometry looking at basophil and mast cell activation following
 127 stimulation with peanut allergen or controls. The basophil activation test (BAT) and the mast
 128 cell activation test (MAT) have high specificity to diagnose peanut allergy, ranging between 96-
 129 100% in different cohorts^{3, 10, 11}; however, they require flow cytometry and fresh blood samples
 130 in the case of BAT. We have previously demonstrated in a peanut allergy model that functional
 131 characteristics of IgE are important in determining basophil and mast cell activation in
 132 response to allergens. These functional characteristics include: specificity, specific activity (i.e.
 133 ration of allergen to total IgE), avidity and diversity. Other studies have characterised allergic
 134 diseases, such as asthma, by mapping interactions of sIgE to clusters/pairs of components
 135 identified using ISAC^{12, 13}.

136 In this study, we aimed to apply these principles to improve the diagnosis of peanut allergy
 137 using serology. By using data from the ISAC assay for each individual (n=100), we built
 138 diagnostic models based upon allergen sensitisation across a broad range of allergens (n=112)
 139 to better diagnose peanut allergy. Initial diagnostic models were developed to incorporate
 140 different input variables, such as ISAC data to a range of allergen components from different

sources, peanut allergen components alone, singleplex peanut allergen component sIgE and the functional characteristics of IgE, that we have recently demonstrated are important in determining whether allergen can induce mast cell and basophil activation in IgE sensitized children.

Methods:

Study population:

Samples from participants in the “Diagnostic markers of clinical allergy versus sensitization to peanut” were tested. This included peanut allergic (PA), peanut sensitized tolerant (PS) and peanut non-sensitized non-allergic (NA) children and teenagers assessed for possible peanut allergy at a specialized Allergy clinic, as previously described³. Peanut allergy was defined by positive oral food challenge (OFC) or the combination of a clear clinical history of systemic allergic reactions to peanut and evidence of IgE sensitization on skin prick test (SPT) of 8 mm or greater and/or peanut-sIgE of 15 kU_A/L or greater¹⁴. Peanut tolerance was defined by a negative OFC or the ability to ingest 4 g or more of peanut protein, twice a week, without developing any allergic symptoms¹⁵. Peanut sensitization was determined by an SPT wheal size of 1 mm or greater and/or a peanut-sIgE level of 0.1 kU_A/L or greater. One hundred patients were randomly selected for the primary population and 60 additional patients for the external validation population, based upon sufficient sample volume. Ethics approval was obtained from the South East London Research Ethics Committee 2, and written informed consent was obtained from parents of all children.

Measurement of IgE specificity, specific activity, avidity and diversity:

Allergen-specific IgE to 112 different allergens was measured using ImmunoCAP ISAC. Total IgE, specific IgE to allergen extracts and specific IgE to individual peanut components (Ara h 1, 2, 3, 6, 8 and 9) were also measured using ImmunoCAP assays. The assays were performed according to the manufacturer’s instructions, as previously reported⁵. IgE specific activity was calculated from the ratio of allergen specific IgE to total IgE. IgE avidity for peanut was measured with an inhibition assay with sodium thiocyanate (NaSCN) using the ImmunoCAP platform and IgE diversity was calculated from the results of IgE binding to peanut allergens on the ISAC test, as described in a previous publication¹⁶.

Statistical analyses:

Four diagnostic models were created: model 1 using IgE to the 112 allergens on the ISAC microarray; model 2 using IgE to 6 peanut components on the ISAC microarray; model 3 using IgE to 4 peanut components on the singleplex ImmunoCAP; and model 4 using four characteristics of peanut-specific IgE: levels, specific activity, avidity and diversity. Statistical modelling was performed employing a generalised linear model (GLM) in a Bayesian hierarchical framework^{17, 18}, using the PA and PS individuals (n=90) as training set. For internal and external validation, the NA subjects were also considered, making totals of 100 and 60 patients respectively. The GLM assumes a Bernoulli distribution for the response variable (PA=1 and PS=0), the input variables are the allergen/characteristic. Regularization is employed to reduce overfitting by using an adaptive prior for the variance hyperparameter, resulting in shrinkage of coefficients for input variables with little influence towards 0.

For diagnostic model 1, the input variables included ImmunoCAP ISAC data. Due to the sparse and high dimensional nature of the data, i.e. number of input variables larger than available samples, sensitisation was considered binary. The input values for each allergen were converted to binary (0/1) format based on the condition that if the detection of allergen sIgE <0.3 then set to 0, otherwise 1. For diagnostic models 2, 3 and 4, raw input values were used instead due to small number of input variables compared to sample size. As the data was not transformed into a binary format for models 2 and 3, we do see outliers appear in this model, hence, a more robust version of the GLM previously described, was used in order to tolerate outliers. This is achieved by modifications to present the model as a finite mixture model, where one component of the mixture is a hierarchical logistic regression model, and the second component is a random 'guessing' parameter adjusting for outliers¹⁷. This provides a 'weighted' version of the estimated coefficients, reducing the unwanted effects of outliers in the data. The script for each model is available online using the link [https://github.com/uhkniazi/BRC Allergy Oliver PID 18/tree/Combining Allergen Components Paper](https://github.com/uhkniazi/BRC%20Allergy%20Oliver%20PID%2018/tree/Combining%20Allergen%20Components%20Paper).

To assess the diagnostic performance, Receiver Operating Characteristic (ROC) curve analyses were performed. In each model, internal 10-fold cross-validation (CV), was used to validate and assess the diagnostic performance of the initial training data (n=90)¹⁹. Where available, diagnostic models 1 and 2, an external validation cohort (n=60; PA=29, PS=24 and NA=7) was used to independently assess the performance of the diagnostic model.

Statistical modelling was performed using both R²⁰ and Stan^{21, 22}, with ROC curves generated in 'R'. Graphical representations of coefficients for each allergen/characteristic were generated using GraphPad Prism 7.0 (GraphPad Software, San Diego, CA).

Results:

Study population

The clinical and demographic characteristics of the primary study population and the cohort used for external validation are represented in **Table 1**.

Diagnostic Model 1 - using the 112 allergen components of ISAC microarray

Diagnostic Model 1 utilised data collected for IgE sensitisation to 112 allergens included in the ISAC microarray. Regression coefficients for the highest and lowest scoring 10 allergens are represented in **Figure 1A** and a full list of allergen regression coefficients in **Table E1**. Ara h 2, Ara h 6 and Ara h 3 reported the highest regression coefficients, suggesting that they hold the most influence in predicting peanut allergy. Ara h 2 reported the highest regression coefficient of 3.05 (SE=1.6), closely followed by Ara h 6 (3.00, SE=1.64) and Ara h 3 (2.33, SE=2.38). Of the peanut components, Ara h 2 reported the lowest standard error alongside the highest coefficient, suggesting that it offers both the greatest reliability with the greatest predictive power. The lowest regression coefficient was exhibited by Phl p 12 (-1.69, SE=2.23). The other peanut allergens, Ara h 1, Ara h 9 and Ara h 8 reported coefficients of 0.203 (SE=1.33), 0.0533 (SE=1.69) and -0.524 (SE=1.58).

The diagnostic performance of model 1 was assessed using 2 different methods. Firstly, we used 10-fold cross-validation as a method of internal validation, which generated an average AUC of 0.82 (**Figure 1B**). Secondly, we used a method of external validation, in which we applied the model to an independent external validation cohort (n=60). The diagnostic performance in the external validation cohort had a greater area under the ROC curve (AUC=0.92). This improvement, despite the external validation ROC curve largely falling within the error bars of the CV curve, suggests that initial CV was particularly rigorous. This improvement of diagnostic performance in the external validation cohort, also overcomes initial concerns that the classifier may have been overtrained in the initial cohort.

Diagnostic Model 2 - using 6 peanut allergen components of ISAC microarray

Diagnostic Model 2 utilised data collected for IgE sensitisation to the 6 peanut allergen components included on the ImmunoCAP ISAC and, as before, was built with data from both PA (n=50) and PS (n=40) individuals. The regression coefficients for peanut allergen components are represented in **Figure 2A**. Allergen sIgE to Ara h 2 and Ara h 6 report the higher regression coefficients of 4.410 (SE=2.620) and 3.090 (SE=2.190). Levels of sIgE to Ara h 2 report the highest predictive power with the greatest reliability, followed by Ara h 6. Ara h 3 reported the 3rd highest regression coefficient of 0.846 (SE=1.560). Ara h 8, Ara h 1 and Ara h 9, all reported negative regression coefficients of -0.525 (SE=0.552), -1.170 (SE=1.080) and -3.00 (SE=2.460), respectively.

As with diagnostic model 1, the diagnostic performance of diagnostic model 2 was assessed using 2 different methods. Internal 10-fold cross-validation of the training data generated an AUC of 0.87 and has been represented by the black curve with error bars in **Figure 2B**. The diagnostic performance in the independent external validation cohort (n=60), was slightly worse than the diagnostic performance seen in the training data, reporting an AUC of 0.86. The similarities between the two reported AUC's suggest that the classifier was not overtrained and performed similarly with internal validation as with external validation.

Diagnostic Model 3 - using 4 peanut allergen components of singleplex ImmunoCAP

Diagnostic Model 3 utilised data collected for IgE sensitisation to the 4 major peanut allergen components Ara h 1, 2, 3 and 6 using the singleplex ImmunoCAP platform, across both PA (n=50) and PS (n=40) individuals. Each of the regression coefficients for peanut allergen components have been represented in **Figure 3A**. Similar to diagnostic models 1 and 2, Ara h 2-sIgE, followed by Ara h 6-sIgE, reported the greatest regression coefficients of 7.2 (SE=3.58) and 4.83 (SE=2.48), respectively. Ara h 1 and Ara h 3 both reported negative coefficients, albeit having lesser influence on the model than sIgE to Ara h 2 and Ara h 6, which reported regression coefficients of -2 (SE=1.46) and -3.62 (SE=2.63), respectively.

As shown in **Figure 3B**, combination of sIgE to Ara h 1, Ara h 2, Ara h 3 and Ara h 6 reported a large AUC of 0.94, when tested in the initial cohort (n=100). The diagnostic performance within the internal cross-validation reports a similar AUC of 0.92. Notably, the error bars from the AUC for the internal cross-validation encompass the AUC performed on the initial cohort

(n=100), suggesting the model had not been overtrained although this would need to be confirmed by assessing the diagnostic performance in an external validation cohort.

Diagnostic Model 4 - using functional characteristics of IgE

Diagnostic Model 4 was developed using data generated on the functional characteristics of IgE: peanut-sIgE levels, peanut specific activity, IgE avidity to peanut and IgE diversity to peanut allergens. As data for IgE avidity to peanut was limited to 66 individuals (PA= 39, PS=27) for technical reasons, two models were created: model 4A, built using data from n=90 and including levels, specific activity and diversity; and model 4B, incorporating data from n=66 and including levels, specific activity, diversity and avidity.

Regression coefficients for the functional characteristics of IgE are represented in **Figures 4A and 5A**. For diagnostic model 4A, peanut diversity reported the greatest regression coefficient of 2.83 (SE=0.913), followed by peanut specific activity with a coefficient of 0.376 (SE=0.159) and peanut-sIgE with a coefficient of 0.00538 (SE=0.00798). For diagnostic model 4B, the coefficients are as follows: peanut diversity (0.609, SE=0.907), peanut specific activity (0.281, SE=0.207), peanut avidity (0.106, SE=0.0354) and peanut-sIgE (0.00538, SE=0.00798).

Considering the diagnostic performance of models 4A and 4B, the two ROC curves report very similar AUC's in both the initial cohort (n=100) and the internal 10-fold cross-validation (n=90) – **Figure 4B and 5B**. The error bars from the AUC for the internal cross-validation encompass the AUC performed on the initial cohort (n=100), suggesting the model hasn't been overtrained; however, this needs to be confirmed with external validation.

Optimal diagnostic cut-offs and online calculators to support clinical decision-making

From the ROC curve analyses, an optimal cut-off defined by the Youden index was defined for each model and the performance of such cut-offs was calculated (**Table 2**). Comparing the models, it is evident that model 1 has the highest accuracy overall, followed by model 3, and that model 3, followed by model 2, has the highest specificity and positive predictive value, which is what is desirable for component testing, allowing confirmation of peanut allergy in patients who are sensitised to peanut.

Online calculators have been generated and will be made available to allow clinicians to input the results of component testing in order to generate a probability of clinical peanut allergy,

that can be used support clinical decisions on whether to perform peanut OFC and on the management of peanut allergy.

Discussion:

A significant proportion of children are referred for oral peanut challenges every year, which reflects the number of children with an equivocal diagnosis following a careful clinical history and results of tests for allergen-specific IgE, using SPT and/or specific IgE assays. Even with Ara h 2-specific IgE being a highly discriminative marker of peanut allergy, low or absent levels and an ambiguous clinical history can make the assessment of individual cases impossible without performing an OFC. In this study, we aimed to optimise the information provided by peanut IgE serology, which is an inexpensive, automated and standardized test, accessible in most centres. We built diagnostic models of allergen-specific IgE to individual molecules using both multiplex and singleplex assays, as well as functional characteristics of peanut-specific IgE, including levels, specific activity, diversity and avidity to predict peanut allergy. We confirmed that Ara h 2-sIgE and Ara h 6-sIgE are the most discriminative serologic markers when used in isolation and found that the combination of the four major components Ara h 1, 2, 3 and 6 can be superior to the individual markers. Similarly, combining the functional characteristics of IgE is superior to using the levels of peanut-sIgE alone. These diagnostic models can be applied in real-time during clinical consultations using online calculators.

The diagnostic models were generated, firstly, to assess the utility of considering a whole range of 112 allergen components, many with known cross reactivity and common co-sensitisation with peanut allergens, such as PR-10 proteins and lipid transfer proteins, in the diagnosis of peanut allergy. Diagnostic model 1 reported an AUC of 0.82 in the training dataset and when validated in an external validation cohort reported an impressive AUC of 0.92. This ROC curve reports slightly higher AUC's than we previously reported for Ara h 2 (0.916) and Ara h 6 (0.908) sIgE⁵. Looking at the highest and lowest reported regression coefficients, Ara h 2, Ara h 6 and Ara h 3 were the most important allergens in differentiating between PA and PS, which is not surprising but is consistent with previous findings⁵. Ara h 1 did not report a greater regression coefficient in this cohort, which is likely reflective of the large proportion of peanut tolerant individuals that are sensitised to this allergen. Perhaps if this model was not designed to interpret sensitisation in a binary manner, Ara h 1 would have reported a greater regression coefficient. This does not appear to be the case with other components known to cross-react

with Ara h 1, such as walnut vicilin, Jug r 2, which suggests its relative importance predicting peanut allergy in this model²³. It is possible that IgE from allergic individuals is more prone to binding to related allergens with structural homology, than IgE of sensitised but tolerant individuals, however this was not reported in the initial PCA clustering analyses (data not shown). Unsurprisingly, Bet v 1, a major birch pollen allergen cross-reactive with other PR-10 proteins, reported a high positive regression coefficient, alongside Aln g 1 from the PR-10 family²⁴. Both these proteins are cross-reactive with peanut allergen Ara h 8, and this maybe an artefact of that cross-reactivity. Considering the allergen components reporting negative regression coefficients, most interesting is perhaps the profilin Phl p 12, which could give rise to peanut-specific IgE not associated with clinical reactivity. Phl p 12 is also indicative of cross-reactivity between pollen from birch, grass, olive, and *Parietaria* and pollen-food syndrome²⁵,²⁶, possibly explaining its negative influence on the prediction of peanut allergy²⁷.

In diagnostic model 2, the ISAC data was employed using peanut allergen components only. This assessed the importance of peanut diversity and its ability to discriminate between allergic and tolerant children and consequent influence on effector cell activation, as previously reported by us¹⁶. Unfortunately, diagnostic model 2 did not improve the previously reported diagnostic performance of peanut diversity (0.898) or diagnostic model 1. This model is also not superior to using singleplex single component sIgE measurements, such as Ara h 2-sIgE (AUC=0.916) or Ara h 6-sIgE (AUC=0.908). This poorer performance may be a consequence of the semi-quantitative nature of results the ISAC and the nature of the model, in which the lack of conversion to a binary format, allows outliers to become prominent, despite adaptations to the model that make it more robust.

Diagnostic model 3, using peanut allergen components Ara h 1, 2, 3 and 6 measured by singleplex ImmunoCAP, reported the highest AUC across all models (AUC=0.94), highlighting the potential of this approach. Importantly, the AUC of the combination of Ara h 1,2,3 and 6 showed a larger AUC than Ara h 2-sIgE or Ara h 6-sIgE; however external validation is needed to confirm these findings. With model 3 reporting similar AUCs to Ara h 2/6 specific activity, the cost-effectiveness of both approaches must be contemplated, as specific activity requires two singleplex ImmunoCAP tests and diagnostic model 3 requires four singleplex ImmunoCAP tests. In order to be financially viable, at least 5 or more allergens would need to be included in ISAC models, as with fewer tests, it would be cheaper to pursue sIgE levels to four allergens using singleplex technology (diagnostic model 3) or two tests to calculate specific activity.

We previously reported that the functional characteristics of IgE, including specific activity, diversity and avidity, in addition to IgE titers, contributed to the ability of IgE to induce basophil and mast cell activation following stimulation with the peanut allergens¹⁶. In diagnostic model 4, we incorporated peanut-sIgE titer, specific activity, diversity and avidity, and showed that the AUC was 0.93 and 0.92, with and without the inclusion of peanut-sIgE avidity, respectively. The slight difference in AUC's between the two groups, is likely reflective of the change of cohort (larger and more comprehensive in terms of range of peanut-sIgE levels in the latter) rather than the addition of IgE avidity. Model 4 reiterates the importance of IgE function in their ability to differentiate between allergic and sensitised but tolerant individuals, especially when taken in combination rather than in isolation¹⁶. However, like diagnostic model 3, these findings require confirmation by external validation.

Apart from the absence of external validation for models 3 and 4, additional limitations of this study are: the relatively small number of patients in the study (n=160 in total); the reduced number of patients with avidity data reducing the size of model 4B; and the absence of Ara h 6-specific IgE levels using the singleplex ImmunoCAP technology, which could potentially improve model 3; however, we have previously demonstrated that Ara h 6-specific IgE does not seem to add much diagnostic value to Ara h 2-specific IgE. Another limitation is the difference in demographics, namely age, between the primary population and the external validation population, with the latter being younger and therefore maybe having fewer sensitization to airborne allergens. Ideally, these populations would have similar demographics but this happened by chance in the consecutive recruiting method used in the study.

Applying the optimal cut-off defined for each model provides better comparison of the diagnostic performance of the models. Given that SPT and specific IgE to extracts have high sensitivity but low specificity, a high specificity is a desired performance characteristic for additional tests. The model with the best performance overall is model 3, obtained with results from peanut components tested on the ImmunoCAP platform, as it has the highest PPV and the second highest accuracy. It is also the least expensive test to perform, compared to ISAC or the various IgE characteristics, and that is an additional advantage. Other tests, like the basophil activation test to peanut, have shown a higher accuracy and higher PPV, but the BAT is a test that requires fresh blood and flow cytometry and is not as widely available and as easy to perform as IgE testing. Thus, combining serologic markers is a good alternative, depending on the resources available to the practising clinician.

394

395 **Acknowledgements:**

396 The authors wish to thank Paediatric Allergy Team at the Evelina London Children's Hospital
397 for contributing to the recruitment of participants to the study, and acknowledge support by
398 the UK National Institute for Health Research comprehensive Biomedical Research Centre
399 award to Guy's and St Thomas' National Health Service (NHS) Foundation Trust, in partnership
400 with the King's College London and King's College Hospital NHS Foundation Trust. The authors
401 would like to thank Dr Jonas Lidholm for reviewing the final version of the manuscript and
402 Thermofisher for supplying reagents for specific IgE determination through a collaboration
403 agreement with King's College London.

References:

1. Nicolaou N, Poorafshar M, Murray C, Simpson A, Winell H, Kerry G, et al. Allergy or tolerance in children sensitized to peanut: prevalence and differentiation using component-resolved diagnostics. *J Allergy Clin Immunol* 2010; 125:191-7.e1-13.
2. Osborne NJ, Koplin JJ, Martin PE, Gurrin LC, Lowe AJ, Matheson MC, et al. Prevalence of challenge-proven IgE-mediated food allergy using population-based sampling and predetermined challenge criteria in infants. *J Allergy Clin Immunol* 2011; 127:668-76 e1-2.
3. Santos AF, Douiri A, Becares 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.
4. Dang TD, Tang M, Choo S, Licciardi PV, Koplin JJ, Martin PE, et al. Increasing the accuracy of peanut allergy diagnosis by using Ara h 2. *J Allergy Clin Immunol* 2012; 129:1056-63.
5. 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.
6. Keet C, Plesa M, Szelag D, Shreffler W, Wood R, Dunlop J, et al. Ara h 2-specific IgE is superior to whole peanut extract-based serology or skin prick test for diagnosis of peanut allergy in infancy. *J Allergy Clin Immunol* 2021; 147:977-83 e2.
7. Melioli G, Bonifazi F, Bonini S, Maggi E, Mussap M, Passalacqua G, et al. The ImmunoCAP ISAC molecular allergology approach in adult multi-sensitized Italian patients with respiratory symptoms. *Clinical Biochemistry* 2011; 44:1005-11.
8. Gadisseur R, Chapelle J-P, Cavalier E. A new tool in the field of in-vitro diagnosis of allergy: preliminary results in the comparison of ImmunoCAP® 250 with the ImmunoCAP® ISAC. *Clinical Chemistry and Laboratory Medicine (CCLM)* 2011; 49:277-80.
9. Foong RX, Dantzer JA, Wood RA, Santos AF. Improving Diagnostic Accuracy in Food Allergy. *J Allergy Clin Immunol Pract* 2021; 9:71-80.
10. Santos AF, Du Toit G, O'Rourke C, Becares N, Couto-Francisco N, Radulovic S, et al. Biomarkers of severity and threshold of allergic reactions during oral peanut challenges. *J Allergy Clin Immunol* 2020; 146:344-55.
11. Santos AF, Couto-Francisco N, Becares 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 e9.
12. Fontanella S, Frainay C, Murray CS, Simpson A, Custovic A. Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A cross-sectional analysis within a population-based birth cohort. *PLOS Medicine* 2018; 15:e1002691.
13. Roberts G, Fontanella S, Selby A, Howard R, Filippi S, Hedlin G, et al. Connectivity patterns between multiple allergen specific IgE antibodies and their association with severe asthma. *Journal of Allergy and Clinical Immunology* 2020; 146:821-30.
14. Roberts G, Lack G. Diagnosing peanut allergy with skin prick and specific IgE testing. *J Allergy Clin Immunol* 2005; 115:1291-6.
15. Fox AT, Sasieni P, du Toit G, Syed H, Lack G. Household peanut consumption as a risk factor for the development of peanut allergy. *Journal of Allergy and Clinical Immunology* 2009; 123:417-23.
16. Hemmings O, Niazi U, Kwok M, James LK, Lack G, Santos AF. Peanut IgE diversity and specific activity are the main determinants of effector cell activation. *J Allergy Clin Immunol* 2020.
17. Kruschke JK. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. 2nd ed; 2014.
18. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. Third ed; 2013.
19. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. 1 ed. New York: Springer; 2013.
20. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing

; 2018.] Available from <https://www.R-project.org/>.

21. Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, et al. Stan: A probabilistic programming language. *Journal of Statistical Software* 76 2017.
22. RStan: the R interface to Stan. 2018.] Available from <http://mc-stan.org>.
23. Barre A, Sordet C, Culierrier R, Rancé F, Didier A, Rougé P. Vicilin allergens of peanut and tree nuts (walnut, hazelnut and cashew nut) share structurally related IgE-binding epitopes. *Molecular Immunology* 2008; 45:1231-40.
24. Wensing M, Akkerdaas JH, van Leeuwen WA, Stapel SO, Bruijnzeel-Koomen CAFM, Aalberse RC, et al. IgE to Bet v 1 and profilin: Cross-reactivity patterns and clinical relevance. *Journal of Allergy and Clinical Immunology* 2002; 110:435-42.
25. Lund G, Brand S, Ramos T, Jimeno L, Boissy P, Vega F, et al. Strong and frequent T-cell responses to the minor allergen Phl p 12 in Spanish patients IgE-sensitized to Profilins. *Allergy* 2018; 73:1013-21.
26. Alvarado MI, Jimeno L, De La Torre F, Boissy P, Rivas B, Lázaro MJ, et al. Profilin as a severe food allergen in allergic patients overexposed to grass pollen. *Allergy* 2014; 69:1610-6.
27. Valenta R, Twaroch T, Swoboda I. Component-resolved diagnosis to optimize allergen-specific immunotherapy in the Mediterranean area. *J Investig Allergol Clin Immunol* 2007; 17:36-40.

Tables:**Table 1.** Demographic and clinical characteristics of the study population.

Characteristics	Primary population			External validation population		
	Peanut Allergic (n=50)	Peanut Sensitised but Tolerant (n=40)	Non Allergic (n=10)	Peanut Allergic (n=29)	Peanut Sensitised but Tolerant (n=24)	Non Allergic (n=6)
Age	8.3 (1.68-17.91)	5.9 (0.52-15.81)	6.0 (0.77-12.58)	6.5 (0.7; 15.0)	1.8 (0.5; 10.3)	11.6 (5.3; 13.6)
Males	36 (72%)	25 (62.5%)	8 (80%)	12 (41%)	14 (48%)	4 (14%)
Total IgE (kU _A /L)	297.5 (4-3550)	469 (7-10714)	29.5 (6-397)	336 (3; 1750)	65 (6; 8042)	217 (52; 580)
Specific IgE to peanut (kU _A /L)	12 (0.20-568)	2.41 (0.04-128)	0.01 (0.01-0.90)	13.6 (0.19; 411)	2.07 (0.05; 22.2)	0.02 (0.01; 0.08)
Specific IgE to Ara h 1 (kU _A /L)	0.80 (0.01-199)	0.11 (0.01-88.3)	0 (0.01-0.01)	3.26 (0.01; 87.6)	0.14 (0.01; 11.7)	0.01 (0.01; 0.03)
Specific IgE to Ara h 2 (kU _A /L)	2.54 (0.01-278)	0.07 (0.01-82.3)	0.04 (0.01-0.07)	8.46 (0.03; 400)	0.07 (0.01; 0.59)	0.04 (0.01; 0.08)
Specific IgE to Ara h 3 (kU _A /L)	0.36 (0.01-89.6)	0.06 (0.01-7.28)	0.01 (0.01-0.04)	0.12 (0.01; 67.9)	0.09 (0.01; 3.28)	0.02 (0.02; 0.04)
Specific IgE to Ara h 6 (kU _A /L)	3.84 (0.01-155)	0.03 (0.01-86)	0 (0.01-0.02)	NP	NP	NP
Specific IgE to Ara h 8 (kU _A /L)	0.09 (0.01-185)	0.04 (0.01-88.6)	0 (0.01-0.02)	0.08 (0.01; 45.2)	0.01 (0.01-27.1)	0.01 (0.01; 0.01)
Specific IgE to Ara h 9 (kU _A /L)	0.02 (0.01-871)	0.05 (0.01-43.9)	0.01 (0.01-0.02)	0.03 (0.01; 4.76)	0.01 (0.01; 1.71)	0.01 (0.01; 0.01)

Footnote: Values are expressed median (range) or as no. (%).

Number of participants tested:

PA: tIgE-18, Peanut-sIgE-23, Ara h 1 sIgE-23, Ara h 2 sIgE -24, Ara h 3 sIgE – 23, Ara h 8 sIgE-22, Ara h 9 sIgE- 22;

PS: tIgE-23, Peanut-sIgE-24, Ara h 1 sIgE-24, Ara h 2 sIgE -24, Ara h 3 sIgE – 24, Ara h 8 sIgE-24, Ara h 9 sIgE- 24;

NA: tIgE-7, Peanut-sIgE-7, Ara h 1 sIgE-7, Ara h 2 sIgE -7, Ara h 3 sIgE -7, Ara h 8 sIgE-7, Ara h 9 sIgE-7.

Table 2. Cut-offs for the various diagnostic models and their performance to diagnose peanut allergy.

Model	Cut-off	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	LR+	LR-	Accuracy (%)
1	0.569	92.4 (84.2; 97.2)	93.8 (84.8; 98.3)	94.8 (87.5; 97.9)	90.9 (82.4; 95.6)	14.8 (5.7; 38.3)	0.08 (0.04; 0.18)	93.0 (82.4; 95.6)
2	0.757	82.3 (72.1; 90.0)	96.9 (89.2; 99.6)	97.0 (89.1; 99.2)	81.7 (73.5; 87.8)	26.3 (6.7; 103.4)	0.18 (0.11; 0.29)	88.9 (73.5; 87.8)
3	0.566	86.0 (73.3; 94.2)	97.5 (86.8; 99.9)	97.7 (85.8; 99.7)	85.1 (74.1; 91.9)	34.4 (5.0; 239.0)	0.14 (0.07; 0.29)	91.2 (74.1; 91.9)
4a	0.484	88.0 (75.7; 95.5)	85.0 (70.2; 94.3)	87.8 (77.3; 93.8)	85.3 (73.0; 92.6)	5.9 (2.8; 12.4)	0.14 (0.07; 0.30)	86.7 (77.9; 92.9)
4b	0.477	87.2 (72.6; 95.7)	85.2 (66.3; 95.8)	89.4 (77.3; 95.5)	82.2 (66.7; 91.4)	5.88 (2.36; 14.66)	0.15 (0.07; 0.35)	86.4 (75.7; 93.6)

Footnote: PPV= Positive Predictive Value, NPV= Negative Predictive Value, LR+= Positive likelihood ratio, LR-= Negative likelihood ratio. 95% Confidence intervals are reported alongside values.

Figure legends:

Figure 1A. Regression coefficients and standard error of the top and bottom 10 allergens for the prediction of peanut allergy in diagnostic model 1. Coefficients are represented by orange bars, with the standard errors overlaid.

Figure 1B. Receiver operator characteristic (ROC) curves for Diagnostic Model 1. The ROC Curve generated on the external validation cohort (n=60) is displayed as a red dashed line. The ROC Curve generated using internal 10-fold cross-validation on the training data (n=90), is displayed as a black line w/ error bars.

Figure 2A. Regression coefficients and standard error of allergen included in diagnostic model 2 for the prediction of peanut allergy. Coefficients are represented by orange bars, with the standard errors overlaid.

Figure 2B. Receiver operator characteristic (ROC) curves for Diagnostic Model 2. The ROC Curve generated on the external validation cohort (n=60) is displayed as a red dashed line. The ROC Curve generated using internal 10-fold cross-validation on the training data (n=90), is displayed as a black line with error bars.

Figure 3A. Regression coefficients and standard error of allergen sIgE levels included in diagnostic model 3 for the prediction of peanut allergy.

Figure 3B. Receiver operator characteristic (ROC) curves for Diagnostic Model 3. The ROC Curve generated on the initial training data cohort; with the addition of NA individuals (n=100) is displayed as a red dashed line. The ROC Curve generated using internal 10-fold cross-validation on the training data (n=90), is displayed as a black line with error bars.

Figure 4A. Regression coefficients and standard errors of Functional IgE characteristics included in diagnostic model 4A for the prediction of peanut allergy. Coefficients are represented by orange bars, with the standard errors overlaid.

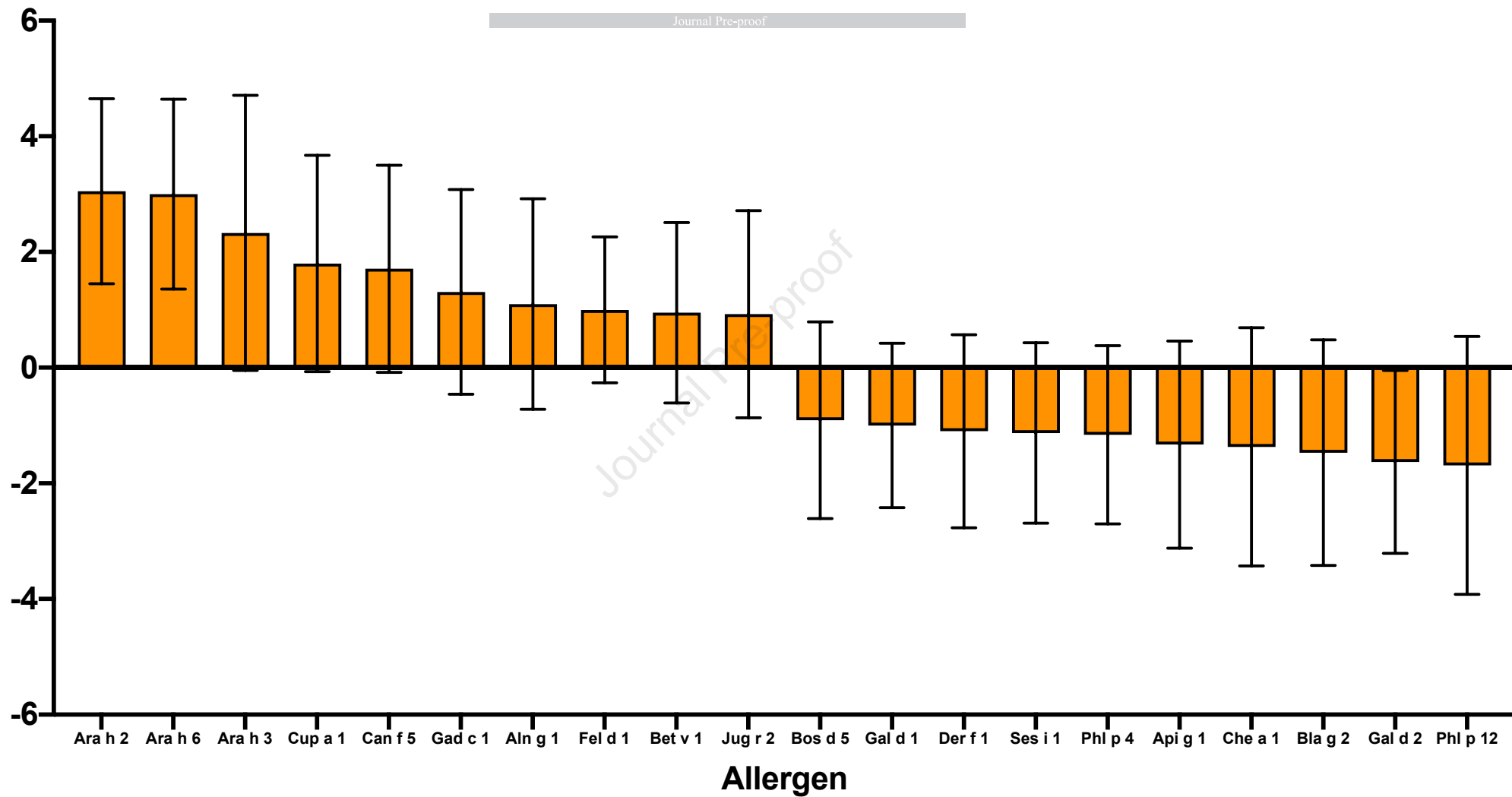
Figure 4B. Receiver operator characteristic (ROC) curves for Diagnostic Model 4A. The ROC curve generated on the initial training data cohort; with the addition of NA individuals (n=100) is displayed as a red dashed line. The ROC Curve generated using internal 10-fold cross-validation on the training data (n=90), is displayed as a black line with error bars.

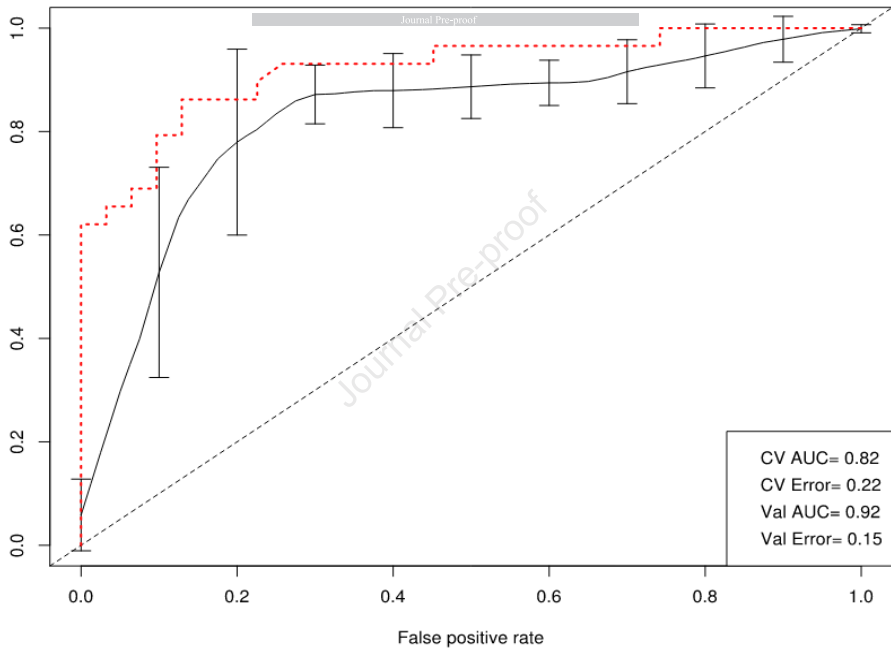
Figure 5A. Regression coefficients and standard errors of Functional IgE characteristics included in diagnostic model 4B for the prediction of peanut allergy. Coefficients are represented by orange bars, with the standard errors overlaid.

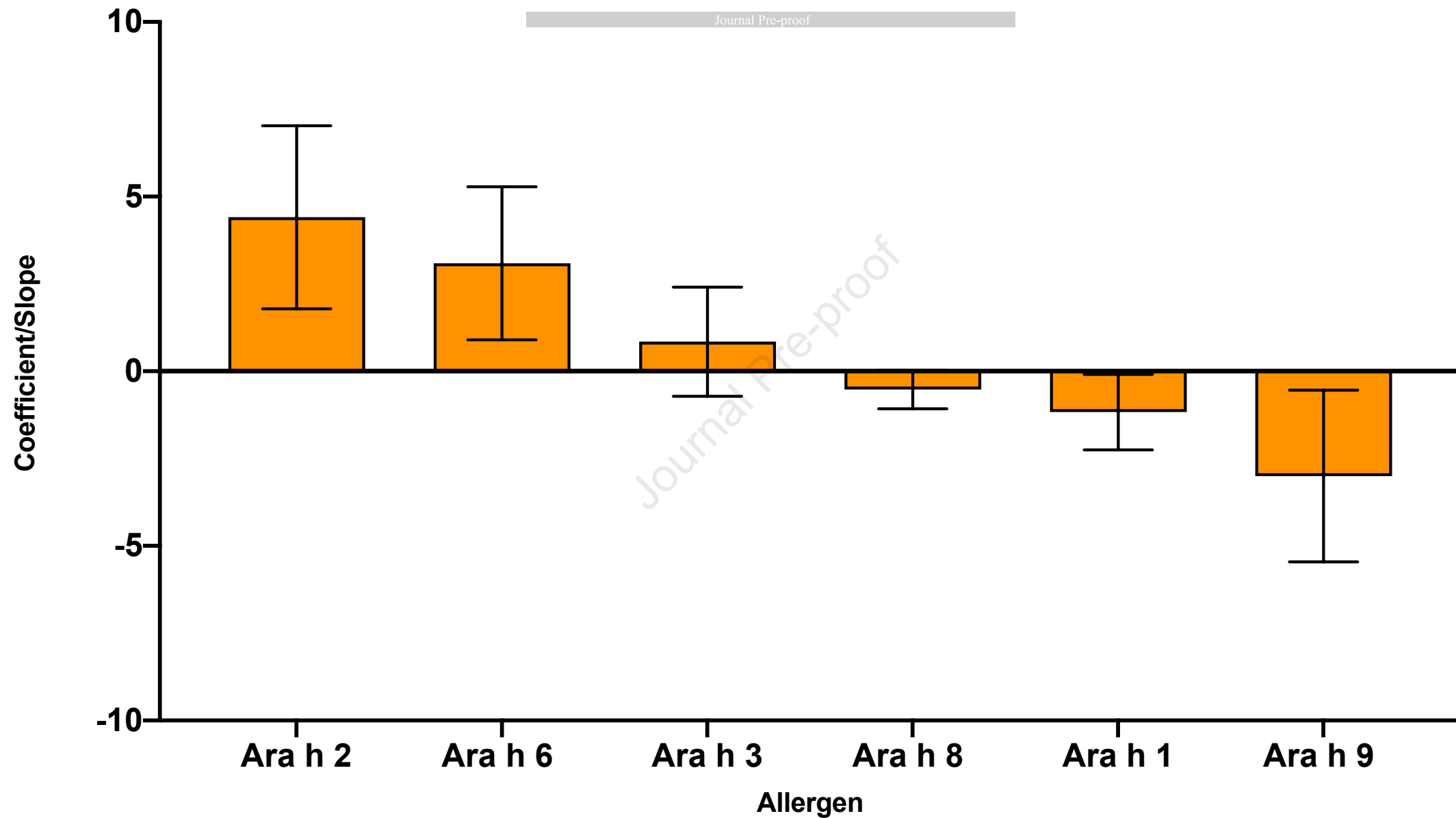
Figure 5B. Receiver operator characteristic (ROC) curves for Diagnostic Model 4B. The ROC Curve generated on the initial training data cohort (n=66) is displayed as a red dashed line. The ROC Curve generated using internal 10-fold cross-validation on the training data (n=66), is displayed as a black line with error bars.

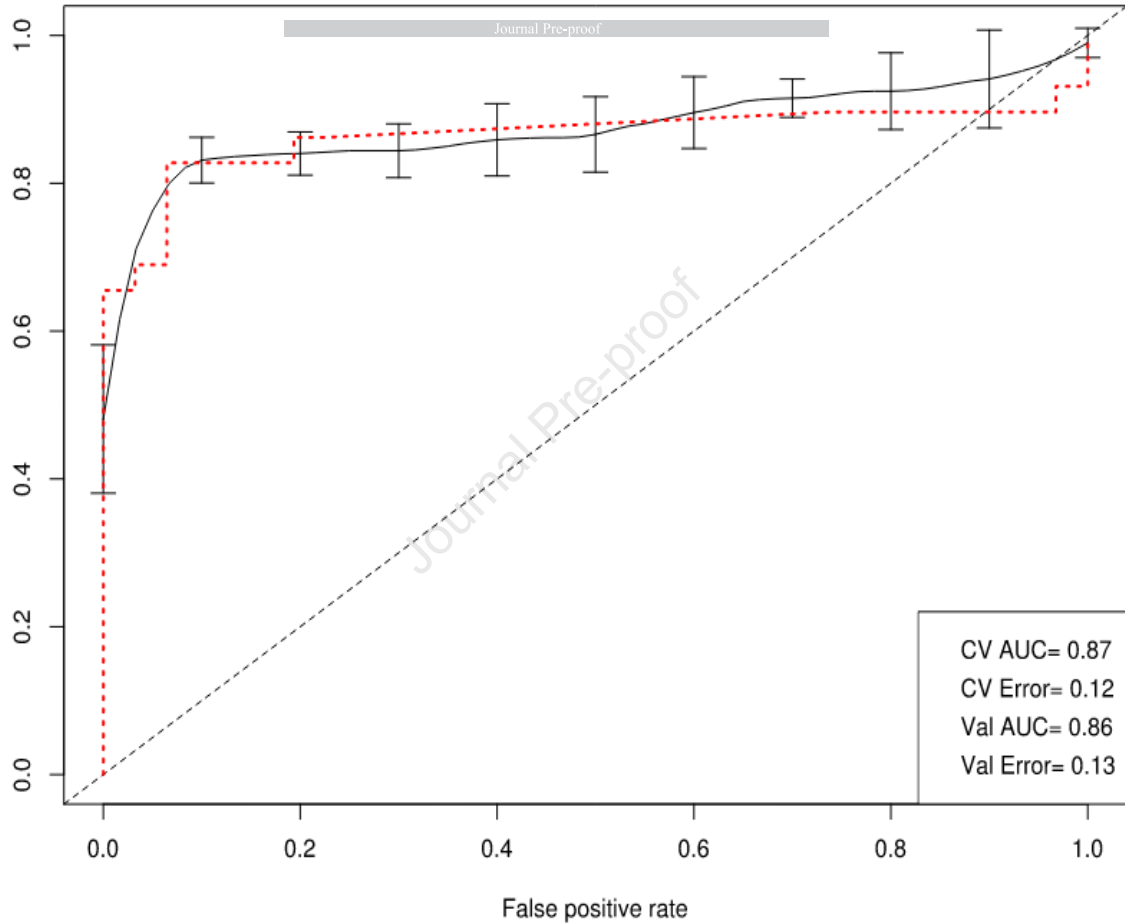
Journal Pre-proof

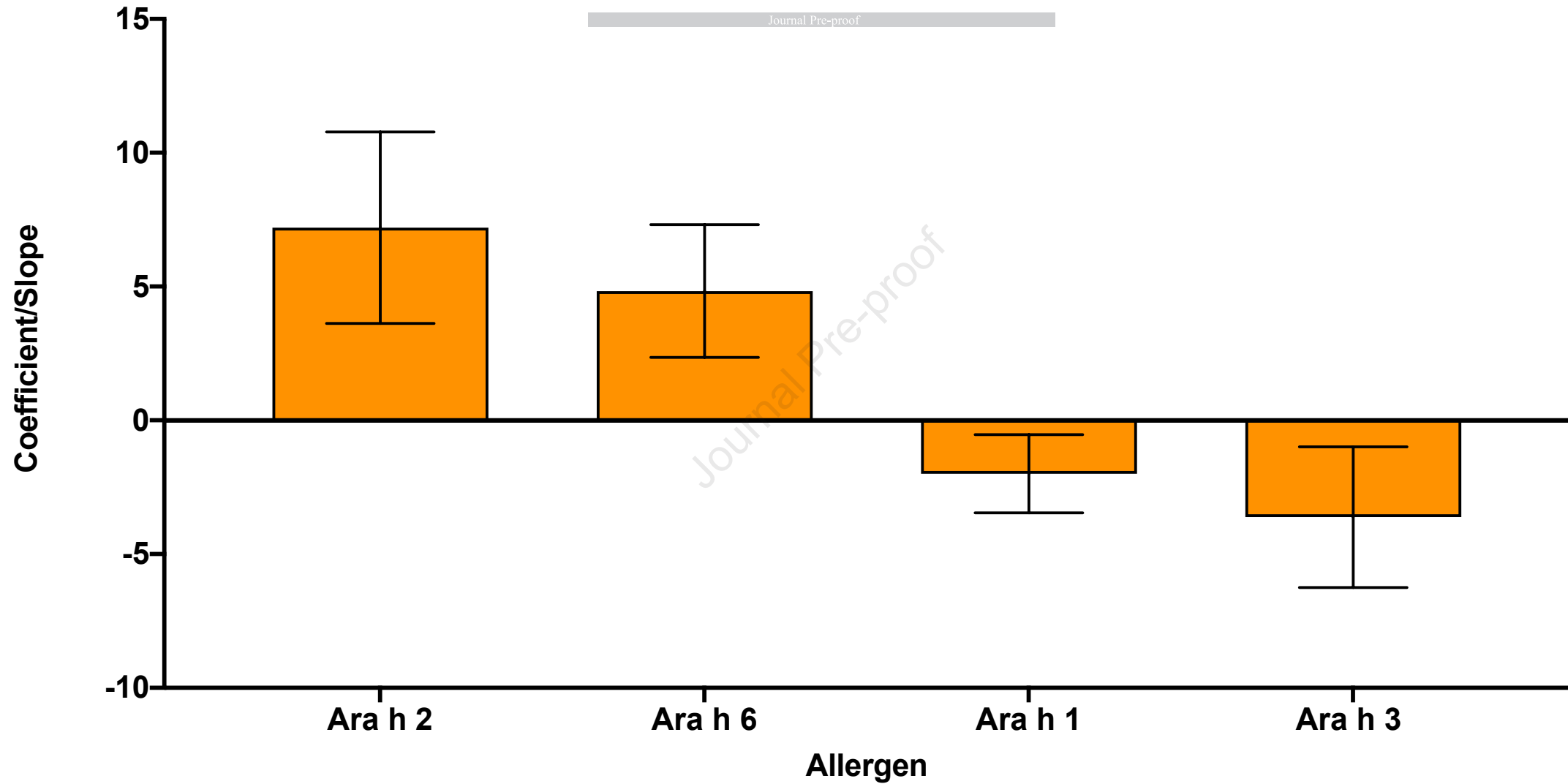
Coefficient/Slope

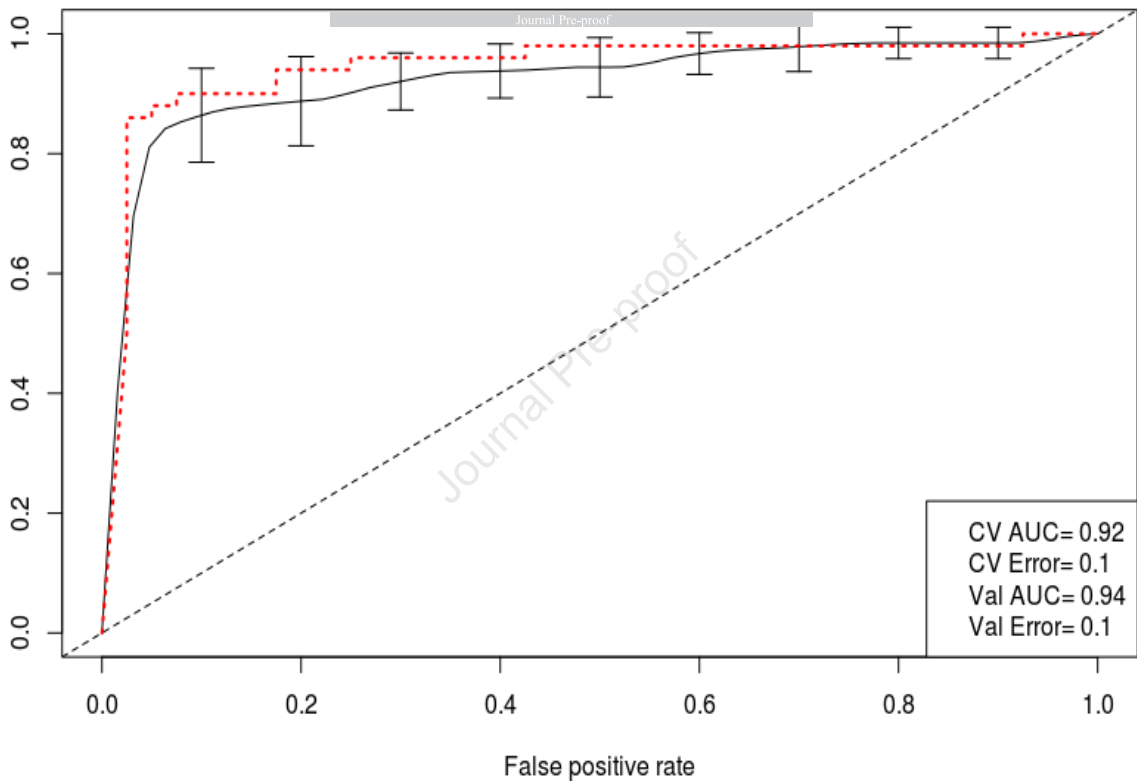




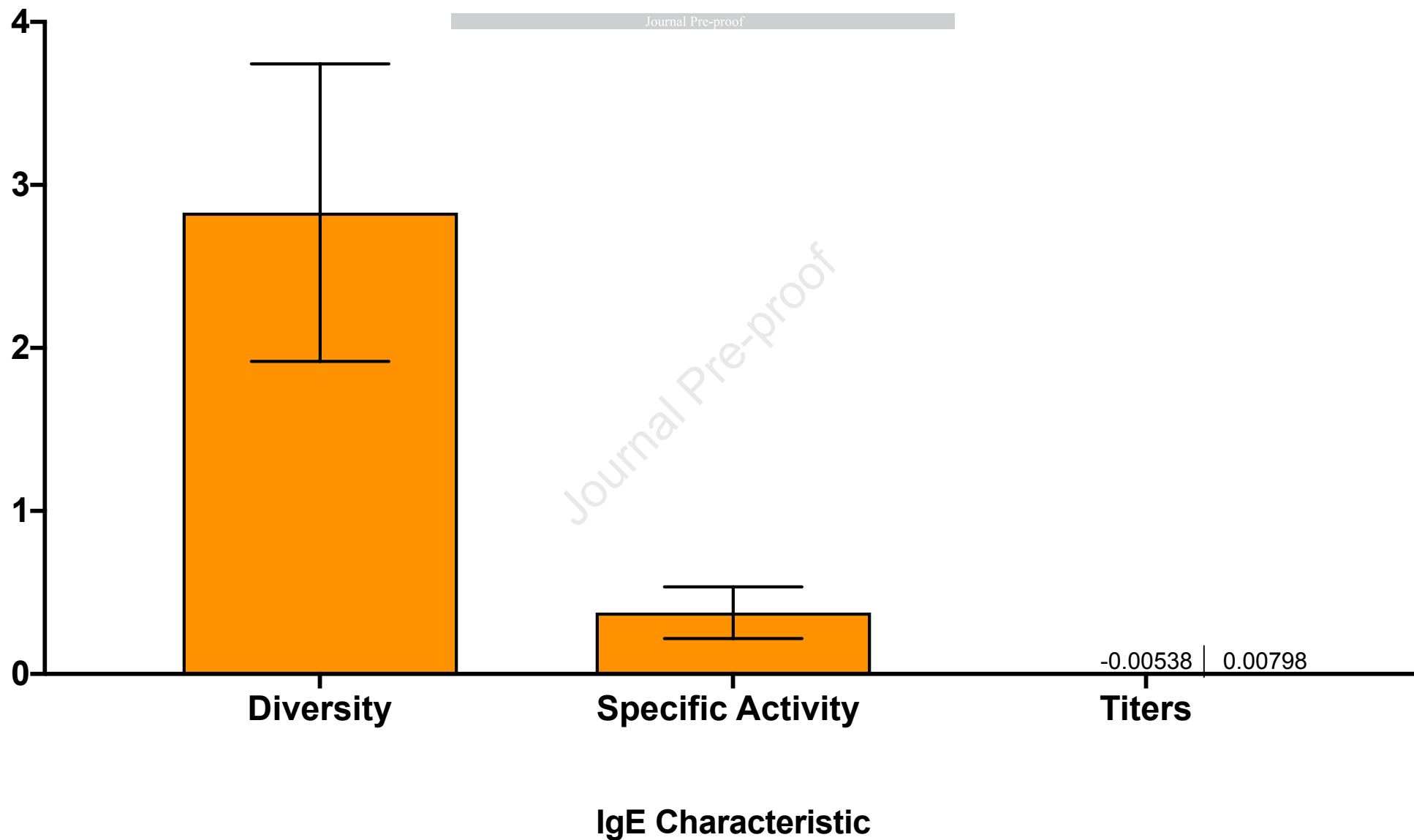


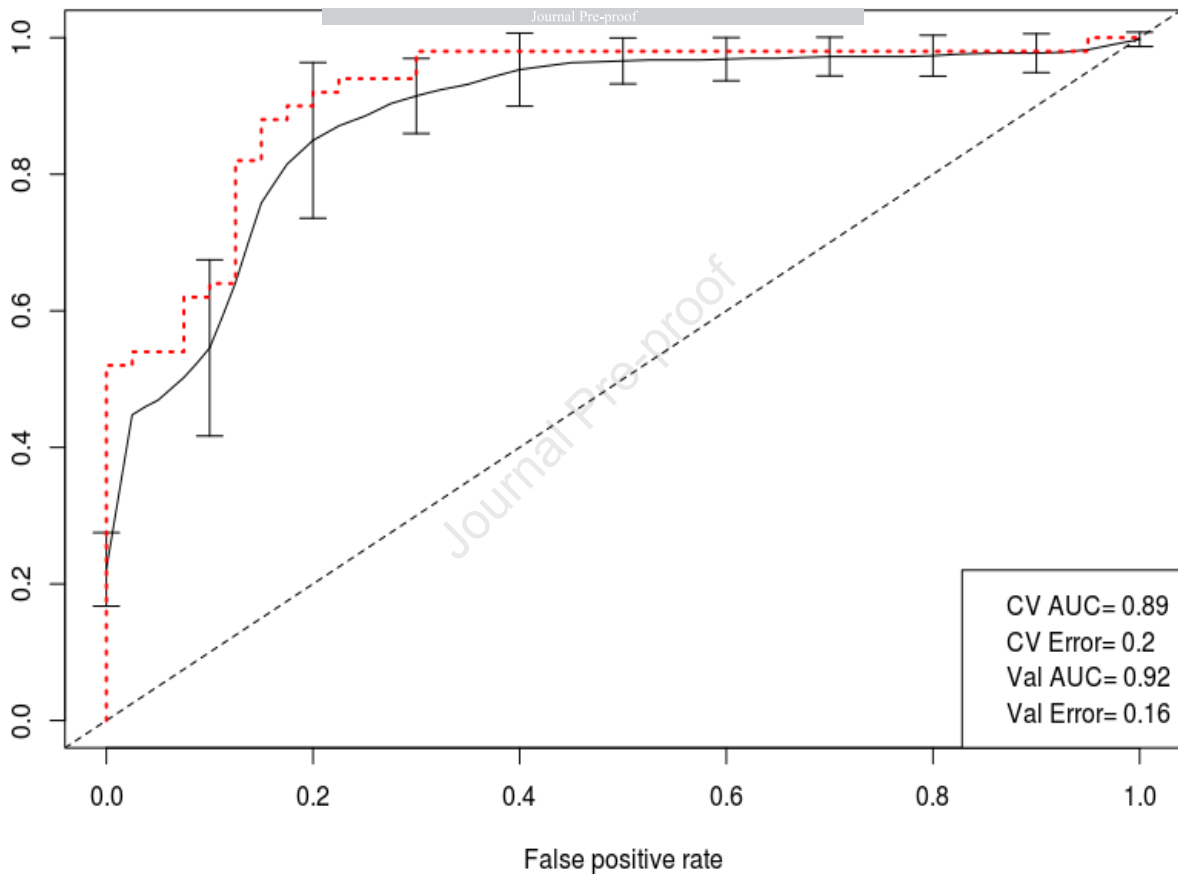


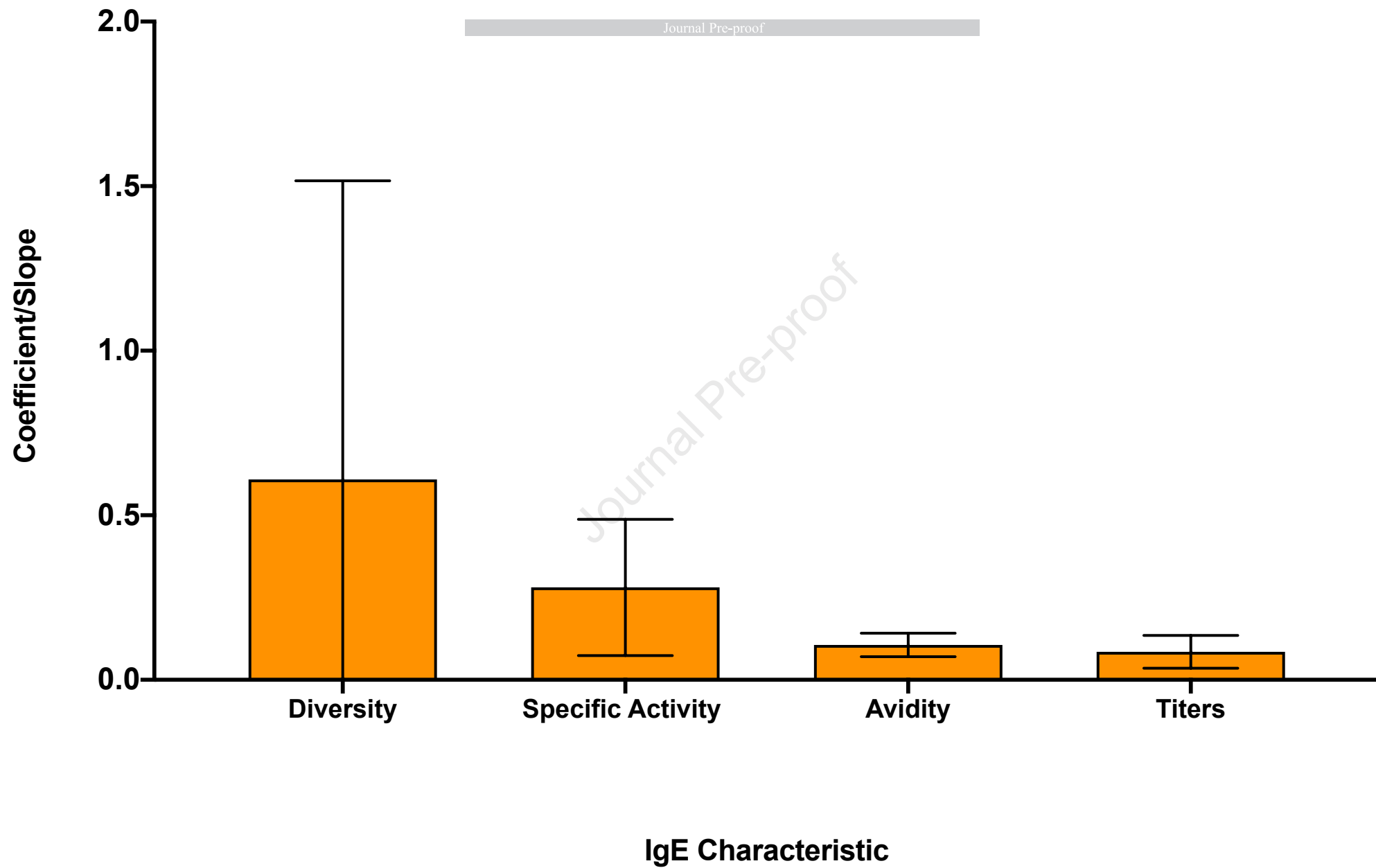


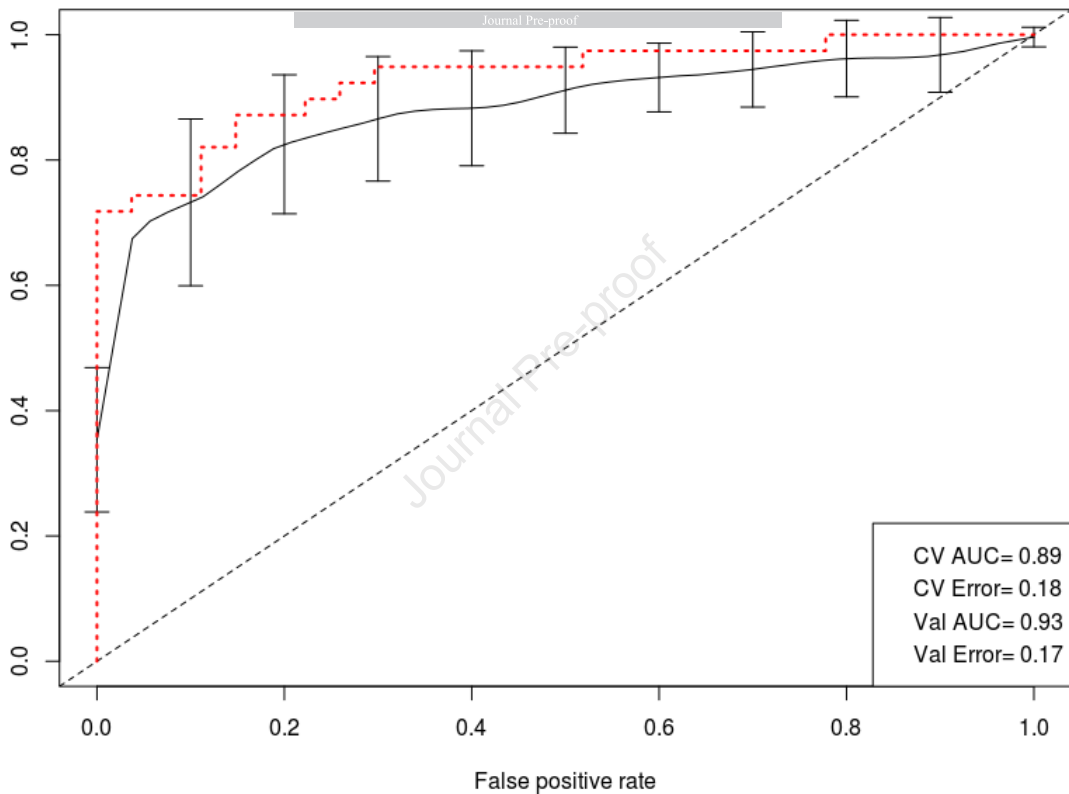


Coefficient/Slope









Online Repository

Combining allergen components improves the accuracy of peanut allergy diagnosis

Oliver Hemmings, MSc^{1,2,3}; Umar Niazi, MBBS PhD⁴; Matthew Kwok, BSc^{1,2,3}; Suzana Radulovic, MD^{1,2,6}; George Du Toit, MBBS FRCPCH^{1,2,6}; Gideon Lack, MBBS FRCPCH^{1,2,3,6}; Alexandra F. Santos, MD PhD^{1,2,3,6*}

¹Department of Women and Children's Health (Pediatric Allergy), School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom

²Peter Gorer Department of Immunobiology, School of Immunology and Microbial Sciences, King's College London, London, United Kingdom

³Asthma UK Centre in Allergic Mechanisms of Asthma, London, United Kingdom

⁴Guy's and St Thomas' National Health Service Foundation Trust and King's College London National Institute for Health Research Biomedical Research Centre Translational Bioinformatics Platform, Guy's Hospital, London, UK.

⁵Blizard Institute, Queen Mary University of London, London, UK

⁶Children's Allergy Service, Evelina London Children's Hospital, Guy's and St Thomas' Hospital, London, United Kingdom

***Corresponding Author:**

Alexandra F. Santos

Address: Department of Paediatric Allergy, 2nd floor, South Wing, St Thomas' Hospital, SE1 7EH London, United Kingdom

Telephone number: +44 (0) 20 7188 6424

Fax number: +44 (0) 20 7403 8640

Email address: alexandra.santos@kcl.ac.uk

E-Methods

Script for Diagnostic models

Models were built using ‘Stan’, and data formatting, preparation and running of the models was performed using R. The script from both Stan and R has been reported below, in that order.

1. Diagnostic model 1

1.1. Stan Script

```
data
{
    int<lower=1> Ntotal; // number of observations
    int<lower=1> Ncol; // total number of columns in model matrix
    matrix[Ntotal, Ncol] X; // model matrix
    int y[Ntotal]; // response variable binomial distributed
}
// transformed data {
// }
parameters {
    // parameters to estimate in the model
    vector[Ncol] betas; // regression parameters
    //vector<lower=0>[(Ncol-1)] sigmas; // scale parameters
    real<lower=0> tau; // standard deviation for deflections
    real<lower=0, upper=1> guess;
}
transformed parameters {
    vector[Ntotal] mu; // fitted values from linear predictor
    vector[Ncol] betas2;
    betas2[1] = betas[1];
    // tau can be close to zero hence parameterized this way
    betas2[2:Ncol] = betas[2:Ncol] * tau;
    mu = X * betas2;
    mu = 0.5 * guess + (1-guess) * inv_logit(mu);
}
model {
    //sigmas ~ cauchy(0, 1);
    tau ~ exponential(1);
    guess ~ beta(1,9);
    betas[1] ~ cauchy(0, 10); //prior for the betas
    betas[2:Ncol] ~ normal(0, 1); //

    // likelihood function
    y ~ bernoulli(mu);
}
generated quantities {
    int y_new[Ntotal];
```

```

vector[Ntotal] log_lik;
y_new = bernoulli_rng(mu);
for (i in 1:Ntotal) log_lik[i] = bernoulli_lpmf(y[i] | mu[i]);
}

```

1.2. R Script

Name:

01_modelAndCV.R

```

# Auth: umar.niazi@kcl.ac.uk
# Date: 13/10/2020
# Desc: model fitting with cross validation and roc curves

```

```

##### data loading and cleaning
source('header.R')

```

```

## load the training set
df = read.csv(file.choose(), header=T)
dim(df)

```

```

# remove NA patients
f = is.na(df$Allergic.Status)
table(f)
df = df[!f,]
dim(df)
# remove any additi
df = na.omit(df)
dim(df)
## remove white space
colnames(df)
df$Sample = factor(gsub(' ', '', as.character(df$Sample)))
df$Allergic.Status = factor(gsub(' ', '',
as.character(df$Allergic.Status)), levels = c('PS', 'PA'))
df = droplevels.data.frame(df)

```

```

## make count matrix
mData = as.matrix(df[, -c(1:2)])
dfSample = df[, 1:2]
rownames(mData) = as.character(dfSample$Sample)
str(dfSample)

```

```
## remove NAs and 0s and convert other values to 1 by adding a jitter
f = mData < 0.3
table(f)
mData[f] = 0
mData[!f] = 1
dim(na.omit(mData))
dim(mData)
```

```
lData.train = list(data=mData, covariates=dfSample)
rm(df)
```

```
#### load the test/validation data
df = read.csv(file.choose(), header=T)
dim(df)
```

```
# remove NA patients
f = is.na(df$Allergic.Status)
table(f)
df = df[!f,]
dim(df)
# remove any additi
df = na.omit(df)
dim(df)
## remove white space
colnames(df)
df$Sample = factor(gsub(' ', '', as.character(df$Sample)))
df$Allergic.Status = factor(gsub(' ', '',
as.character(df$Allergic.Status)), levels = c('PS', 'PA'))
df = droplevels.data.frame(df)
```

```
## make count matrix
mData = as.matrix(df[, -c(1:2)])
dfSample = df[, 1:2]
rownames(mData) = as.character(dfSample$Sample)
str(dfSample)
```

```
## remove NAs and 0s and convert other values to 1 by adding a jitter
f = mData < 0.3
table(f)
mData[f] = 0
mData[!f] = 1
```

```

dim(na.omit(mData))
dim(mData)

lData.test = list(data=mData, covariates=dfSample)
rm(df)
rm(mData)
## sanity check
table(colnames(lData.train$data) %in% colnames(lData.test$data))
identical(colnames(lData.train$data), colnames(lData.test$data))
##### end data loading

##### Stan section for binomial regression approach
### fit 2 models one standard binomial and one with robust version
dfData = rbind(data.frame(lData.train$data),
data.frame(lData.test$data))
dim(dfData)
dfData$fGroups =
factor(c(as.character(lData.train$covariates$Allergic.Status),

as.character(lData.test$covariates$Allergic.Status)), levels = c('PS',
'PA'))
lData = list(resp=ifelse(dfData$fGroups == 'PA', 1, 0),
mModMatrix=model.matrix(fGroups ~ 1 + ., data=dfData))

```

2. Diagnostic model 2

2.1. Stan Script

```

data
{
    int<lower=1> Ntotal; // number of observations
    int<lower=1> Ncol; // total number of columns in model matrix
    matrix[Ntotal, Ncol] X; // model matrix
    int y[Ntotal]; // response variable binomial distributed
}
// transformed data {
// }
parameters {
    // parameters to estimate in the model
    vector[Ncol] betas; // regression parameters
    //vector<lower=0>[(Ncol-1)] sigmas; // scale parameters
    real<lower=0> tau; // standard deviation for deflections
    real<lower=0, upper=1> guess;
}
transformed parameters {
    vector[Ntotal] mu; // fitted values from linear predictor
    vector[Ncol] betas2;
    betas2[1] = betas[1];
    // tau can be close to zero hence parameterized this way
    betas2[2:Ncol] = betas[2:Ncol] * tau;
    mu = X * betas2;
    mu = 0.5 * guess + (1-guess) * inv_logit(mu);
}
model {
    //sigmas ~ cauchy(0, 1);
    tau ~ exponential(1);
    guess ~ beta(1,9);
    betas[1] ~ cauchy(0, 10); //prior for the betas
    betas[2:Ncol] ~ normal(0, 1); //

    // likelihood function
    y ~ bernoulli(mu);
}

```

2.2. R Script

Name:

01_modelAndCV.R

```

# Auth: umar.niazi@kcl.ac.uk
# Date: 13/10/2020

```

```

# Desc: model fitting with cross validation and roc curves

##### data loading and cleaning
source('header.R')

## load the training set
df = read.csv(file.choose(), header=T)
dim(df)

# remove NA patients
f = is.na(df$Allergic.Status)
table(f)
df = df[!f,]
dim(df)
# remove any additi
df = na.omit(df)
dim(df)
## remove white space
colnames(df)
df$Sample = factor(gsub(' ', '', as.character(df$Sample)))
df$Allergic.Status = factor(gsub(' ', '',
as.character(df$Allergic.Status)), levels = c('PS', 'PA'))
df = droplevels.data.frame(df)

## make count matrix
mData = as.matrix(df[, -c(1:2)])
dfSample = df[, 1:2]
rownames(mData) = as.character(dfSample$Sample)
str(dfSample)

## remove NAs and 0s and convert other values to 1 by adding a jitter
# f = mData < 0.3
# table(f)
# mData[f] = 0
# mData[!f] = 1
dim(na.omit(mData))
dim(mData)

lData.train = list(data=mData, covariates=dfSample)
rm(df)

```

```

#### load the test/validation data
df = read.csv(file.choose(), header=T)
dim(df)

# remove NA patients
f = is.na(df$Allergic.Status)
table(f)
df = df[!f,]
dim(df)
# remove any additi
df = na.omit(df)
dim(df)
## remove white space
colnames(df)
df$Sample = factor(gsub(' ', '', as.character(df$Sample)))
df$Allergic.Status = factor(gsub(' ', '',
as.character(df$Allergic.Status)), levels = c('PS', 'PA'))
df = droplevels.data.frame(df)

## make count matrix
mData = as.matrix(df[, -c(1:2)])
dfSample = df[, 1:2]
rownames(mData) = as.character(dfSample$Sample)
str(dfSample)

mData = mData[, colnames(lData.train$data)]
## remove NAs and 0s and convert other values to 1 by adding a jitter
# f = mData < 0.3
# table(f)
# mData[f] = 0
# mData[!f] = 1
dim(na.omit(mData))
dim(mData)

lData.test = list(data=mData, covariates=dfSample)
rm(df)
rm(mData)
## sanity check
table(colnames(lData.train$data) %in% colnames(lData.test$data))
identical(colnames(lData.train$data), colnames(lData.test$data))

```

```
##### end data loading
```

```
##### Stan section for binomial regression approach
```

```
### fit 2 models one standard binomial and one with robust version
```

```
dfData = rbind(data.frame(lData.train$data),
```

```
data.frame(lData.test$data))
```

```
dim(dfData)
```

```
dfData$fGroups =
```

```
factor(c(as.character(lData.train$covariates$Allergic.Status),
```

```
as.character(lData.test$covariates$Allergic.Status)), levels = c('PS',  
'PA'))
```

```
lData = list(resp=ifelse(dfData$fGroups == 'PA', 1, 0),
```

```
mModMatrix=model.matrix(fGroups ~ 1 + ., data=dfData))
```

3. Diagnostic model 3

3.1. Stan Script

```

data {
  int<lower=1> Ntotal; // number of observations
  int<lower=1> Ncol; // total number of columns in model matrix
  matrix[Ntotal, Ncol] X; // model matrix
  int y[Ntotal]; // response variable binomial distributed
}
// transformed data {
// }
parameters {
  // parameters to estimate in the model
  vector[Ncol] betas; // regression parameters
  //vector<lower=0>[(Ncol-1)] sigmas; // scale parameters
  real<lower=0> tau; // standard deviation for deflections
  real<lower=0, upper=1> guess;
}
transformed parameters {
  vector[Ntotal] mu; // fitted values from linear predictor
  vector[Ncol] betas2;
  betas2[1] = betas[1];
  // tau can be close to zero hence parameterized this way
  betas2[2:Ncol] = betas[2:Ncol] * tau;
  mu = X * betas2;
  mu = 0.5 * guess + (1-guess) * inv_logit(mu);
}
model {
  //sigmas ~ cauchy(0, 1);
  tau ~ exponential(1);
  guess ~ beta(1,9);
  betas[1] ~ cauchy(0, 10); //prior for the betas
  betas[2:Ncol] ~ normal(0, 1); //

  // likelihood function
  y ~ bernoulli(mu);
}

```

3.1.1.1.1. R Script

Name:

01_modelAndCV.R

```

# Auth: umar.niazi@kcl.ac.uk
# Date: 13/10/2020
# Desc: model fitting with cross validation and roc curves

```

```
##### data loading and cleaning
source('header.R')

## load the training set
df = read.csv(file.choose(), header=T)
dim(df)

# remove NA patients
f = is.na(df$Allergic.Status)
table(f)
df = df[!f,]
dim(df)
# remove any additi
df = na.omit(df)
dim(df)
## remove white space
colnames(df)

df$Patient.ID = factor(gsub(' ', '', as.character(df$Patient.ID)))
df$Allergic.Status = factor(gsub(' ', '',
as.character(df$Allergic.Status)), levels = c('PS', 'PA'))
df = droplevels.data.frame(df)

## make count matrix
mData = as.matrix(df[, -c(1:2)])
dfSample = df[, 1:2]
rownames(mData) = as.character(dfSample$Patient.ID)
str(dfSample)

dim(na.omit(mData))
dim(mData)

lData.train = list(data=mData, covariates=dfSample)
rm(df)

##### end data loading
```

```
##### Stan section for binomial regression approach
### fit 2 models one standard binomial and one with robust version
dfData = data.frame(lData.train$data)
dim(dfData)
dfData$fGroups = lData.train$covariates$Allergic.Status
lData = list(resp=ifelse(dfData$fGroups == 'PA', 1, 0),
mModMatrix=model.matrix(fGroups ~ 1 + ., data=dfData))
```

4. Diagnostic model 4

4.1. Stan Script

```

data
{
    int<lower=1> Ntotal; // number of observations
    int<lower=1> Ncol; // total number of columns in model matrix
    matrix[Ntotal, Ncol] X; // model matrix
    int y[Ntotal]; // response variable binomial distributed
}
// transformed data {
// }
parameters {
    // parameters to estimate in the model
    vector[Ncol] betas; // regression parameters
    //vector<lower=0>[(Ncol-1)] sigmas; // scale parameters
    real<lower=0> tau; // standard deviation for deflections
    real<lower=0, upper=1> guess;
}
transformed parameters {
    vector[Ntotal] mu; // fitted values from linear predictor
    vector[Ncol] betas2;
    betas2[1] = betas[1];
    // tau can be close to zero hence parameterized this way
    betas2[2:Ncol] = betas[2:Ncol] * tau;
    mu = X * betas2;
    mu = 0.5 * guess + (1-guess) * inv_logit(mu);
}
model {
    //sigmas ~ cauchy(0, 1);
    tau ~ exponential(1);
    guess ~ beta(1,9);
    betas[1] ~ cauchy(0, 10); //prior for the betas
    betas[2:Ncol] ~ normal(0, 1); //

    // likelihood function
    y ~ bernoulli(mu);
}
generated quantities {
    int y_new[Ntotal];
    vector[Ntotal] log_lik;
    y_new = bernoulli_rng(mu);
    for (i in 1:Ntotal) log_lik[i] = bernoulli_lpmf(y[i] | mu[i]);
}

```

4.2. R Script

```

# Name:
01_modelAndCV.R

# Auth: umar.niazi@kcl.ac.uk
# Date: 13/10/2020
# Desc: model fitting with cross validation and roc curves

##### data loading and cleaning
source('header.R')

## load the training set
df = read.csv(file.choose(), header=T)
dim(df)

# remove NA patients
f = is.na(df$Allergic.Status)
table(f)
df = df[!f,]
dim(df)
# remove any additi
df = na.omit(df)
dim(df)
## remove white space
colnames(df)

df$Patient_ID = factor(gsub(' ', '', as.character(df$Patient_ID)))
df$Allergic.Status = factor(gsub(' ', '',
as.character(df$Allergic.Status)), levels = c('PS', 'PA'))
df = droplevels.data.frame(df)

## make count matrix
mData = as.matrix(df[, -c(1:2)])
dfSample = df[, 1:2]
rownames(mData) = as.character(dfSample$Patient_ID)
str(dfSample)

dim(na.omit(mData))
dim(mData)

```

```
lData.train = list(data=mData, covariates=dfSample)
rm(df)
```

```
##### end data loading
```

```
##### Stan section for binomial regression approach
### fit 2 models one standard binomial and one with robust version
dfData = data.frame(lData.train$data)
dim(dfData)
dfData$fGroups = lData.train$covariates$Allergic.Status
lData = list(resp=ifelse(dfData$fGroups == 'PA', 1, 0),
mModMatrix=model.matrix(fGroups ~ 1 + ., data=dfData))
```

E-Tables

Table E1. List of 112 Allergens included in ImmunoCAP-ISAC microarray with respective coefficients

Allergen Component	Allergen type	Allergen Source	Protein Group	Coefficient (Diagnostic Model 1)	S.E.M. (Diagnostic Model 1)
nGal d 1	Food Allergen	Egg	Ovomucoid	-1	1.42
nGal d 2	Food Allergen	Egg	Ovalbumin	-1.63	1.58
nGal d 3	Food Allergen	Egg	Conalbumin/Ovotransferrin	-0.397	1.74
nGal d 5	Food Allergen	CSA (Livetin)	Serum albumin	-0.528	1.64
nBos d 4	Food Allergen	Cow's milk	Alpha-lactalbumin	0.0856	1.68
nBos d 5	Food Allergen	Cow's milk	Beta-lactoglobulin	-0.91	1.7
nBos d 6	Food Allergen	BSA	Serum albumin	-0.432	1.72
nBos d 8	Food Allergen	Cow's milk	Casein	0.166	1.56
nBos d lactoferrin	Food Allergen	Cow's milk	Transferrin	-0.684	1.89
rGad c 1	Food Allergen	Cod	Parvalbumin	1.31	1.77
nPen m 1	Food Allergen	Shrimp	Tropomyosin	0.74	1.78
nPen m 2	Food Allergen	Shrimp	Arginine kinase	0.0818	1.89
nPen m 4	Food Allergen	Shrimp	Sarcoplasmic Ca-binding protein	-0.313	1.65
rAna o 2	Food Allergen	Cashew nut	Storage protein, 11S globulin	-0.128	1.38
rBer e 1	Food Allergen	Brazil nut	Storage protein, 2S albumin	0.0291	1.5
rCor a 1.0401	Food Allergen	Hazelnut	PR-10 protein	0.448	1.58
rCor a 8	Food Allergen	Hazelnut	Lipid transfer protein (nsLTP)	0.775	1.74
nCor a 9	Food Allergen	Hazelnut	Storage protein, 11S globulin	-0.554	1.45
nJug r 1	Food Allergen	Walnut	Storage protein, 2S albumin	-0.245	1.44
nJug r 2	Food Allergen	Walnut	Storage protein, 7S globulin	0.924	1.79
nJug r 3	Food Allergen	Walnut	Lipid transfer protein (nsLTP)	-0.33	1.67

nSes i 1	Food Allergen	Sesame seed	Storage protein, 2S albumin	-1.13	1.56
rAra h 1	Food Allergen	Peanut	Storage protein ,7S globulin	0.203	1.33
rAra h 2	Food Allergen	Peanut	Storage protein, Conglutin	3.05	1.6
rAra h 3	Food Allergen	Peanut	Storage protein, 11S globulin	2.33	2.38
nAra h 6	Food Allergen	Peanut	Storage protein, Conglutin	3	1.64
rAra h 8	Food Allergen	Peanut	PR-10 protein	-0.524	1.58
rAra h 9	Food Allergen	Peanut	Lipid transfer protein (nsLTP)	0.0533	1.69
rGly m 4	Food Allergen	Soybean	PR-10 protein	-0.638	1.78
nGly m 5	Food Allergen	Soybean	Storage protein, Beta-conglycinin	0.817	1.71
nGly m 6	Food Allergen	Soybean	Storage protein, Glycinin	0.435	1.46
nFag e 2	Food Allergen	Buckwheat	Storage protein, 2S albumin	-0.0322	1.84
rTri a 14	Food Allergen	Wheat	Lipid transfer protein (nsLTP)	-0.0435	1.83
rTri a 19.0101	Food Allergen	Wheat	Crude gliadin	0.0573	1.77
nTri a aA_TI	Food Allergen	Wheat	Omega-5 gliadin	-0.773	1.8
nAct d 1	Food Allergen	Kiwi		-0.759	1.49
nAct d 2	Food Allergen	Kiwi	Thaumatine-like protein	0.12	1.55
nAct d 5	Food Allergen	Kiwi		0.0139	1.89
rAct d 8	Food Allergen	Kiwi	PR-10 protein	-0.25	1.62
rApi g 1	Food Allergen	Celery	PR-10 protein	-1.33	1.79
rMal d 1	Food Allergen	Apple	PR-10 protein	0.733	1.59
rPru p 1	Food Allergen	Peach	PR-10 protein	0.0775	1.67
rPru p 3	Food Allergen	Peach	Lipid transfer protein (nsLTP)	0.161	1.64
nCyn d 1	Aero Allergen	Bermuda grass	Grass group 1	0.023	1.38
rPhl p 1	Aero Allergen	Timothy grass	Grass group 1	-0.0853	1.33
rPhl p 2	Aero Allergen	Timothy grass	Grass group 2	0.00222	1.28

nPhl p 4	Aero Allergen	Timothy grass		-1.16	1.54
rPhl p 5	Aero Allergen	Timothy grass	Grass group 5	-0.0172	1.35
rPhl p 6	Aero Allergen	Timothy grass		-0.808	1.41
rPhl p 7	Aero Allergen	Timothy grass	Polcalcin	0.356	1.89
rPhl p 11	Aero Allergen	Timothy grass		-0.22	1.75
rPhl p 12	Aero Allergen	Timothy grass	Profilin	-1.69	2.23
rAln g 1	Aero Allergen	Alder	PR-10 protein	1.1	1.82
rBet v 1	Aero Allergen	Birch	PR-10 protein	0.949	1.56
rBet v 2	Aero Allergen	Birch	Profilin	-0.262	1.59
rBet v 4	Aero Allergen	Birch	Polcalcin, Calcium bonding protein	0.329	1.88
rCor a 1.0101	Aero Allergen	Hazel pollen	PR-10 protein	0.369	1.65
nCry j 1	Aero Allergen	Japanese cedar		-0.69	1.66
nCup a 1	Aero Allergen	Cypress		1.8	1.87
nOle e 1	Aero Allergen	Olive		-0.524	1.28
nOle e 7	Aero Allergen	Olive	Lipid transfer protein (nsLTP)	-0.296	1.67
rOle e 9	Aero Allergen	Olive		0.696	1.76
rPla a 1	Aero Allergen	Plane tree		-0.265	1.65
nPla a 2	Aero Allergen	Plane tree		0.663	1.51
rPla a 3	Aero Allergen	Plane tree	Lipid transfer protein (nsLTP)	-0.257	1.66
nAmb a 1	Aero Allergen	Ragweed		0.00535	1.79
nArt v 1	Aero Allergen	Mugwort		-0.148	1.73
nArt v 3	Aero Allergen	Mugwort	Lipid transfer protein (nsLTP)	0.0997	1.5
rChe a 1	Aero Allergen	Goosefoot		-1.37	2.06
rMer a 1	Aero Allergen	Annual mercury	Profilin	-0.324	1.58
rPar j 2	Aero Allergen	Wall pellitory	Lipid transfer protein (nsLTP)	-0.0373	1.81

rPla l 1	Aero Allergen	Plantain (English)		-0.328	1.78
nSal k 1	Aero Allergen	Saltwort		0.58	1.69
rCan f 1	Aero Allergen	Dog	Lipocalin	0.642	1.26
rCan f 2	Aero Allergen	Dog	Lipocalin	-0.309	1.5
nCan f 3	Aero Allergen	Dog	Serum albumin	-0.533	1.75
rCan f 5	Aero Allergen	Dog	Arginine esterase	1.71	1.79
rEqu c 1	Aero Allergen	Horse	Lipocalin	-0.881	1.54
nEqu c 3	Aero Allergen	Horse	Serum albumin	0.574	2.02
rFel d 1	Aero Allergen	Cat	Uteroglobin	0.999	1.26
nFel d 2	Aero Allergen	Cat	Serum albumin	-0.308	1.69
rFel d 4	Aero Allergen	Cat	Lipocalin	0.47	1.56
nMus m 1	Aero Allergen	Mouse	Lipocalin	-0.349	1.69
rAlt a 1	Aero Allergen	Alternaria		0.745	1.53
rAlt a 6	Aero Allergen	Alternaria	Enolase	0.0133	1.67
rAsp f 1	Aero Allergen	Aspergillus		-0.0658	1.94
rAsp f 3	Aero Allergen	Aspergillus		0.72	1.91
rAsp f 6	Aero Allergen	Aspergillus	Mn superoxide dismutase	-0.0355	1.78
rCla h 8	Aero Allergen	Cladosporium		0.0865	1.52
rBlo t 5	Aero Allergen	House dust mite		0.2	1.75
nDer f 1	Aero Allergen	House dust mite		-1.1	1.67
rDer f 2	Aero Allergen	House dust mite		0.0159	1.54
nDer p 1	Aero Allergen	House dust mite		-0.438	1.6
rDer p 2	Aero Allergen	House dust mite		0.133	1.61
rDer p 10	Aero Allergen	House dust mite	Tropomyosin	-0.245	1.76
rLep d 2	Aero Allergen	Storage mite		-0.401	1.68

rBla g 1	Aero Allergen	Cockroach		-0.00308	1.89
rBla g 2	Aero Allergen	Cockroach		-1.47	1.95
rBla g 5	Aero Allergen	Cockroach		-0.00513	1.97
nBla g 7	Aero Allergen	Cockroach	Tropomyosin	0.662	1.82
rApi m 1	Other	Honey bee venom	Phospholipase A2	-0.0502	1.7
nApi m 4	Other	Honey bee venom	Melittin	0.0473	1.85
rPol d 5	Other	Paper wasp venom	Venom, Antigen 5	-0.0731	1.8
rVes v 5	Other	Common wasp venom	Venom, Antigen 5	0.0278	1.9
rAni s 1	Other	Anisakis		0.0521	1.85
rAni s 3	Other	Anisakis	Tropomyosin	0.471	1.84
rHev b 1	Other	Latex		-0.00901	1.95
rHev b 3	Other	Latex		0.403	1.83
rHev b 5	Other	Latex		0.0445	1.95
rHev b 6.01	Other	Latex		-0.0166	1.91
rHev b 8	Other	Latex	Profilin	-0.315	1.6
nMUXF3	Other	Sugar epitope from Bromelain	CCD-marker	-0.0266	1.59