

ORIGINAL ARTICLE

Challenges in interpreting allergen microarrays in relation to clinical symptoms: A machine learning approach

Mattia C. F. Prosperi^{1,2}, Danielle Belgrave^{1,2}, Iain Buchan¹, Angela Simpson² & Adnan Custovic²

¹Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK; ²Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester, Manchester, UK

To cite this article: Prosperi MCF, Belgrave D, Buchan I, Simpson A, Custovic A. Challenges in interpreting allergen microarrays in relation to clinical symptoms: A machine learning approach. *Pediatr Allergy Immunol* 2013; **000**.

Keywords

component-resolved diagnostics; asthma; wheeze; rhinitis; airway hyper-reactivity; methacholine; IgE; children; machine learning; feature selection; logistic regression; random forests; Bayesian networks

Correspondence

Mattia C. F. Prosperi, Department of Biomedical Modelling, Centre for Health Informatics, Institute of Population Health, University of Manchester, 1st Floor, Jean McFarlane Building, Room 1.314, University Place, Oxford Road, Manchester M13 9PL, UK
Tel.: +44 (0) 161 27 51125
E-mail: mattia.prosperi@manchester.ac.uk

Accepted for publication 30 August 2013

DOI:10.1111/pai.12139

Abstract

Background: Identifying different patterns of allergens and understanding their predictive ability in relation to asthma and other allergic diseases is crucial for the design of personalized diagnostic tools.

Methods: Allergen-IgE screening using ImmunoCAP ISAC[®] assay was performed at age 11 yrs in children participating a population-based birth cohort. Logistic regression (LR) and nonlinear statistical learning models, including random forests (RF) and Bayesian networks (BN), coupled with feature selection approaches, were used to identify patterns of allergen responses associated with asthma, rhino-conjunctivitis, wheeze, eczema and airway hyper-reactivity (AHR, positive methacholine challenge). Sensitivity/specificity and area under the receiver operating characteristic (AUROC) were used to assess model performance *via* repeated validation.

Results: Serum sample for IgE measurement was obtained from 461 of 822 (56.1%) participants. Two hundred and thirty-eight of 461 (51.6%) children had at least one of 112 allergen components IgE > 0 ISU. The binary threshold >0.3 ISU performed less well than using continuous IgE values, discretizing data or using other data transformations, but not significantly ($p = 0.1$). With the exception of eczema (AUROC~0.5), LR, RF and BN achieved comparable AUROC, ranging from 0.76 to 0.82. Dust mite, pollens and pet allergens were highly associated with asthma, whilst pollens and dust mite with rhino-conjunctivitis. Egg/bovine allergens were associated with eczema.

Conclusions: After validation, LR, RF and BN demonstrated reasonable discrimination ability for asthma, rhino-conjunctivitis, wheeze and AHR, but not for eczema. However, further improvements in threshold ascertainment and/or value transformation for different components, and better interpretation algorithms are needed to fully capitalize on the potential of the technology.

Detection of allergen-specific IgE antibodies (sIgE) is associated with an increased risk of wheeze/asthma, and among asthmatic patients with more severe disease and diminished lung function (1–4). The level of sIgE to common inhalant allergens offers more valuable information than a simple detection of ‘positive sIgE’ (4, 5). Different allergen sources (both indoor and outdoor) have been independently associated with asthma and asthma-related symptoms (6–9). However, it remains unclear how allergen sensitizations *in toto* contribute towards clinical manifestations of different atopic diseases (e.g., asthma vs. rhino-conjunctivitis vs. eczema).

The increasing availability of allergen components (purified from natural source or produced as recombinant proteins) marks the shift in allergy diagnosis that may lead to a

transition towards component-resolved diagnostics (10). For example, the multiplex chip-based assay ImmunoCAP ISAC[®] has been validated in terms of performance and reproducibility (11), providing an opportunity to identify both allergen patterns and their interactions in relation to different clinical outcomes. Using ImmunoCAP ISAC[®] assay, the sIgE antibody profiles associated with asthma, exhaled nitric oxide and airway hyper-reactivity (AHR) have been investigated, with multiple sensitizations to several allergen groups increasing the risk of asthma (12).

We hypothesized that different sIgE patterns are predictive of different diseases commonly associated with atopy; also, different interpretation algorithms, component threshold ascertainment and/or value transformation may modify the

association between such patterns and clinical symptoms. To address these hypotheses, we investigated the ability and the interpretability of different linear and nonlinear statistical learning models in classifying contemporaneous asthma, wheezing, AHR, rhino-conjunctivitis and eczema. Models were fit on sIgE levels measured by the ISAC® among participants in a population-based birth cohort at age 11 yrs. We used machine learning methods such as decision trees (DTs) (that divide study population into nested subgroups, based on allergen thresholds and combinations, each with a specific probability of manifesting clinical symptoms) and Bayesian networks (BN) (graphs that represent causal dependencies of variables), coupled with logistic regression (LR), to fully exploit the large amount of information provided by the microarray, and to associate it to clinical symptoms.

Methods

Study population and data sources

Manchester Asthma and Allergy Study is a population-based birth cohort described in detail elsewhere (13). The study was approved by a local ethics committee; informed consent was obtained from all parents. All data in this manuscript were ascertained at age 11 yrs. We administered validated questionnaires to collect information on parentally reported symptoms and physician-diagnosed illnesses. We measured AHR using methacholine challenge (14).

Definition of outcomes

Current asthma

Positive answer to all three of the following questions: (i) 'Has your child wheezed within the past 12 months?', (ii) 'Has your child received asthma medication within the past 12 months?' and (iii) 'Has your child ever been diagnosed with asthma?'.

Current wheeze

Positive answer to the question 'Has your child had wheezing or whistling in the chest in the last 12 months?'.

Current eczema

Positive answer to the question 'Has your child had eczema in the last 12 months?'.

Current rhino-conjunctivitis

Positive answer to the question 'In the past 12 months, has your child ever had a problem with sneezing, or a runny nose, or a blocked nose when he/she did not have a cold or the flu which was accompanied by itchy-watery eyes?'.

Airway hyper-reactivity

Provocative concentration of methacholine causing a 20% decline in FEV₁ < 16 mg/ml.

All outcomes were encoded as binary variables. Other variables considered in the descriptive statistics were FEV₁, FVC, FEV₁/FVC ratio, eNO and methacholine dose-response ratio.

Detection of IgE antibodies

The presence of sIgE to 112 allergen components was assessed by the ImmunoCAP ISAC® (ThermoFisher Scientific, Uppsala, Sweden).

Transformation of sIgE values

In the analyses described below, we expressed sIgE values as follows: (i) binarized using the threshold of 0.3 ISU; (ii) discretized into four categories using the manufacturer's semiquantitative scale (<0.3 ISU, undetectable or very low; ≥0.3 and <1 ISU, low; ≥1 and <15 ISU, moderate to high; ≥15 ISU, very high); (iii) discretized using an automated supervised discretization approach (15); (iv) continuous raw values; (v) square-root or hyperbolic-arcsine transformation (16); (vi) using other normalization methods such as quantile normalization (17).

Statistical learning

For a detailed explanation of the methods, please see the Supporting Information. Briefly, we analysed the discriminative ability of sIgE patterns in relation to clinical outcomes by fitting a series of machine learning models. All analyses were adjusted for gender. Statistical models were run on the subset of patients with at least one sIgE > 0 ISU.

Logistic regression

We fitted main-effects LR using: (i) the sum of and the number of positive sIgE values; and (ii) sIgE to all allergen components (using different transformations). For the latter, due to the high number of variables, LR was subject to feature selection via LogitBoost (18).

Decision tree and random forest models

These models were fitted to investigate possible nonlinear/interaction effects (19, 20). DTs are machine learning methods that divide population into nested subgroups according to values of the covariates, usually those that have the highest discriminatory power with respect to the outcome. For example, our study population could be divided into two using Fel d 1 IgE below or above 0.3 ISU. Other tree-branching rules can then be inferred on the two subpopulations, and so on recursively until a stopping criterion is met (e.g., a minimum number of subjects per subgroup). This progressive data partition can be represented in the form of a tree (Fig. S1). DTs are easy to interpret, but sometimes have poor predictive power. RFs are an ensemble of several different DTs, fitted with resampling/randomization, with the aim to improve prediction performance by combining many decision pathways (e.g., averaging across many DT predictions). DT and RF have dedicated methods for measuring variable importance which can capture complex interactions, without the need of explicitly defining them.

Bayesian networks

Bayesian networks are graphs in which each node is a covariate, and a link between two nodes represent a dependency. If no link

is present between two nodes, they are conditionally independent. For a comprehensive introduction to BN modelling for biomedicine, please see Millán et al. (21). The naïve Bayes (NB), a simpler model which assumes conditional independence among variables, was also fit as a control to BN.

Model performance

Goodness-of-fit functions for assessing prediction performance of models included as follows: accuracy (% correct), area under the receiver operating characteristic (AUROC), sensitivity and specificity. The ability to generalize on unseen data was assessed through repeated validation, executing for 50 times a randomized training/test procedure (80%/20%) and comparing differences between models with a paired-corrected *t*-test. We also assessed power of the sample in relation to covariate size.

Results

Participants

We reviewed 822 children. Sample for IgE measurement was obtained for 461 (56.1%); there was no difference in gender, family history of allergic diseases, position in sibship, asthma, sensitization (skin tests) or parental atopy between those with

and without IgE (data available on request). A total of 238 of 461 (51.6%) children tested positive (sIgE > 0.3 ISU) to at least one allergen component. Characteristics of study participants are shown in Table 1.

Distribution and transformation of sIgE values

The distribution of sIgE values was highly skewed; Fig. S2 shows the histograms upon several input transformations. In a preliminary test on model performance (AUROC in relation to asthma, using LR and RF), we found that the binary discretization of sIgE values using the 0.3 ISU threshold was performing less well compared with a continuous scale or a multiple categorization. The manufacturer's semiquantitative scale performed better than the binary threshold, although not significantly. The automated supervised discretization method yielded the best results. Fig. S3 shows detailed box plots of AUROC performance across all transformation methods.

Discriminative ability of sIgE patterns in relation to clinical outcomes

Robustness of model performance

Average AUROC > 0.5 was achieved for asthma, wheeze, rhino-conjunctivitis and AHR in all statistical learning models;

Table 1 Characteristics of the study population at age 11 (N = 426)

	Median (IQR) or N			Missing N/A
	Male 154	Female 84	Total 238	
At least one IgE > 0				
N ₁ = 238 (51.6%)				
Number of specific positive IgE (>0.3 ISU)	7.5 (3.0–14.5)	7.0 (2.0–12.0)	7.0 (3.0–13.0)	N/A
Sum of all IgE	34.0 (8.4–130.2)	38.9 (1.9–132.8)	36.4 (4.9–131.4)	N/A
Asthma	33	15	48	4
Eczema	36	22	58	5
Mean eNO	14.6 (8.8–33.8)	19.9 (9.8–41.7)	17.6 (9.4–37.8)	62
FVC	2.7 (2.4–3.0)	2.5 (2.3–2.9)	2.6 (2.4–3.0)	3
% predicted FEV ₁	99.2 (91.0–107.4)	98.5 (92.2–103.7)	99.0 (91.6–106.0)	3
Current wheeze	48	25	73	3
AHR	62	31	93	59
Rhino-conjunctivitis	63	28	91	3
Methacholine dose-response ratio	3.20 (1.01–5.47)	3.48 (0.95–6.11)	0.98 (3.28–5.84)	57
FEV/FVC ratio	0.86 (0.81–0.90)	0.89 (0.84–0.92)	0.86 (0.82–0.91)	4
All IgE = 0				
N ₂ = 223 (48.4%)				
Asthma	5	5	10	9
Eczema	15	14	29	7
Mean eNO	8.3 (6.9–10.7)	8.1 (6.5–10.4)	8.2 (6.6–10.6)	46
FVC	2.7 (2.5–3.1)	2.6 (2.3–2.9)	2.7 (2.4–3.0)	6
% Predicted FEV ₁	98.3 (93.0–106.1)	99.0 (91.8–106.6)	98.6 (92.5–106.5)	5
Current wheeze	8	10	18	2
Airway hyper-reactivity	23	18	41	50
Rhino-conjunctivitis	7	9	16	11
Methacholine dose-response ratio	4.69 (2.88–6.48)	5.13 (3.81–6.81)	4.99 (3.54–6.57)	53
FEV/FVC ratio	0.86 (0.82–0.90)	0.89 (0.85–0.93)	0.88 (0.83–0.92)	13

Table 2 Performance of statistical learning models by means of 50 independent validation runs, stratified by different outcomes

Outcome	Method	Feature set	Feature/topology selection	AUROC (s.d.)	Sensitivity (s.d.)	Specificity (s.d.)
Asthma	Majority class	N/A	N/A	0.50 (0.00)*	1.00 (0.00)	0.00 (0.00)*
	LR	Number of positive IgE + sum of all IgE	N/A	0.71 (0.10)*	0.96 (0.03)	0.20 (0.12)*
	LR	112 IgE + gender	Cross-validated LogitBoost	0.79 (0.08)	0.95 (0.03)*	0.40 (0.15)
	DT	112 IgE + gender	Embedded (information gain, pruning)	0.59 (0.10)*	0.96 (0.06)	0.14 (0.16)
	RF	112 IgE + gender	Embedded (Gini index, random subset)	0.82 (0.06)	0.97 (0.04)	0.34 (0.13)
	NB	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.76 (0.08)	0.91 (0.05)*	0.38 (0.15)
	BN	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.77 (0.07)	0.96 (0.03)*	0.32 (0.15)
Wheeze	Majority class	N/A	N/A	0.50 (0.00)*	1.00 (0.00)	0.00 (0.00)*
	LR	Number of positive IgE + sum of all IgE	N/A	0.67 (0.06)*	0.93 (0.04)*	0.13 (0.08)*
	LR	112 IgE + gender	Cross-validated LogitBoost	0.72 (0.07)	0.94 (0.05)	0.37 (0.12)
	DT	112 IgE + gender	Embedded (information gain, pruning)	0.61 (0.08)*	0.90 (0.09)	0.29 (0.20)
	RF	112 IgE + gender	Embedded (Gini index, random subset)	0.78 (0.06)	0.91 (0.05)*	0.45 (0.12)
	NB	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.69 (0.06)*	0.92 (0.06)*	0.30 (0.12)
	BN	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.65 (0.07)*	0.93 (0.10)	0.29 (0.12)
Rhino-conjunctivitis	Majority class	N/A	N/A	0.50 (0.00)*	1.00 (0.00)	0.00 (0.00)*
	LR	Number of positive IgE + sum of all IgE	N/A	0.80 (0.07)	0.84 (0.07)*	0.44 (0.11)*
	LR	112 IgE + gender	Cross-validated LogitBoost	0.73 (0.07)*	0.79 (0.09)*	0.53 (0.13)
	DT	112 IgE + gender	Embedded (information gain, pruning)	0.66 (0.06)*	0.81 (0.11)*	0.47 (0.17)
	RF	112 IgE + gender	Embedded (Gini index, random subset)	0.78 (0.07)	0.80 (0.08)*	0.57 (0.11)
	NB	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.75 (0.07)	0.88 (0.06)*	0.41 (0.12)*
	BN	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.73 (0.07)*	0.82 (0.08)*	0.50 (0.12)
AHR	Majority class	N/A	N/A	0.50 (0.00)*	1.00 (0.00)	0.00 (0.00)*
	LR	Number of positive IgE + sum of all IgE	N/A	0.57 (0.11)*	0.70 (0.19)*	0.39 (0.20)*
	LR	112 IgE + gender	Cross-validated LogitBoost	0.64 (0.05)*	0.55 (0.18)*	0.70 (0.23)
	DT	112 IgE + gender	Embedded (information gain, pruning)	0.64 (0.10)*	0.55 (0.18)*	0.70 (0.21)
	RF	112 IgE + gender	Embedded (Gini index, random subset)	0.76 (0.07)	0.69 (0.11)*	0.70 (0.10)
	NB	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.64 (0.10)*	0.62 (0.20)*	0.57 (0.25)
	BN	112 IgE + gender	Cross-validated wrapper (best-first search, K2)	0.56 (0.06)*	0.58 (0.21)*	0.54 (0.29)

AHR, airway hyper-reactivity; AUROC, area under the receiver operating characteristic; BN, Bayesian networks; DT, Decision tree; LR, Logistic regression; NB, naïve Bayes; RF, random forests.

*The hypothesis of difference in means comparing against the best model (in bold) could not be rejected at $p = 0.05$.

in contrast, this was not achieved for eczema (AUROC~0.5). However, although AUROC for eczema for all models was poor, in the univariate analysis, sIgE to egg ovomucoid, ovalbumin and ovotransferrin were significantly associated with eczema (Gal d 1, $p = 0.02$; Gal d 2; $p = 0.02$; Gal d 3, $p = 0.02$). We also observed a strong trend for bovine allergens Bos d 4-5-6 ($p = 0.06$, $p = 0.06$, $p = 0.04$, respectively).

Overall, all models showed reasonable AUROC (0.76–0.82 for RF, 0.63–0.79 for LR, 0.56–0.77 for BN, 0.64–0.76 for NB) and sensitivity (0.69–0.97 for RF, 0.54–0.95 for LR, 0.58–0.96 for BN, 0.62–0.92 for NB), but poor specificity (0.34–0.70 for RF, 0.40–0.69 for LR, 0.32–0.54 for BN, 0.38–0.57 for NB), except for the positive AHR (higher specificity, decreased AUROC/sensitivity).

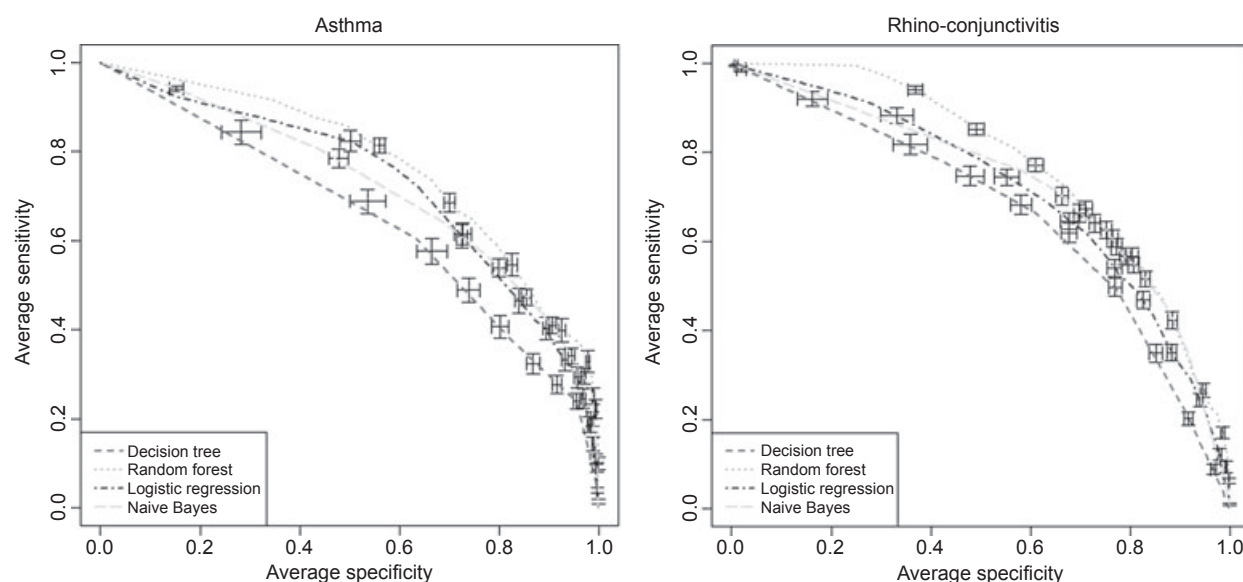


Figure 1 Performance of statistical learning models in classifying the asthma and rhino-conjunctivitis outcomes, using the full feature set (112 IgE + gender) by means of area under the receiver operating characteristic, across 50 independent validation (80%/20%) runs. Results are out-of-sample predictions (i.e., on unseen data). Bars represent standard errors.

Random forest outperformed other approaches in terms of AUROC in all but one outcome (rhino-conjunctivitis). LR (model ii) ranked always as the best in terms of specificity. The number of variables selected by LogitBoost yielded a median (IQR) of 9 (7–18) covariates per model across all validation runs. In most cases, the hypothesis that there was no difference in the mean performance among the RF, LR, NB and BN could not be rejected at the 0.05 level. Performance of DT was consistently inferior to that of RF ($p < 0.05$), and the same held for LR model i (encoding the number of positive IgE + sum of all IgE), except for rhino-conjunctivitis. Table 2 summarizes prediction performance obtained by the repeated validation; AUROC plots are shown in Fig. 1. For this experiment, we used the square-root transformed sIgE, but similar results were obtained for the other transformation methods.

Variable importance in relation to clinical outcomes and their dependencies

Fig. 2 shows RF feature importance plots with respect to asthma and rhino-conjunctivitis across 1000 permutation runs. The importance is expressed as rescaled decrease in accuracy when randomizing a variable of interest. For asthma, there was a broader set of top-scoring allergens belonging to different sources including dust mite, cat, dog and pollens, whilst the top-scoring allergens for rhino-conjunctivitis were all pollens followed by dust mite. The order of variables may change when randomizing permutations (22); indeed, variable ranks and associated p-values for our data set were subject to a consistent degree of variation.

There was partial consistency between the variables selected by the stepwise heuristic for NB/BN with the top-scoring variables output by the RF. Similarly, runs of stepwise LR with different starting points led to different final sets, probably due

to a number of correlated variables. Fig. 3 shows the mutually adjusted odds ratio from LogitBoost LR, including only variables significant ($p < 0.05$) in univariate analysis. Supplementary results give a more thorough explanation of the relevant variables and their association into 'equivalency' groups (Figs S4–S6).

Fig. 4 depicts two optimized NB and BN structures for asthma and rhino-conjunctivitis. The networks shown here are representative of a single run; therefore, we cannot exclude the possibility that there are other topologies and variable sets with comparable performance.

Discussion

Key findings

We investigated the ability of linear and nonlinear statistical learning models fit on ISAC® assay data, to identify asthma, wheezing, AHR, rhino-conjunctivitis and eczema. In general, all modelling techniques (excluding DRs) performed comparably. With the exception of eczema, all outcomes could be predicted with an AUROC > 0.5 . Random forests (RF) outperformed other approaches in terms of AUROC, whilst LR ranked as the best in terms of specificity. We could not clarify if a main-effect model or a model that hypothesizes conditional independence among variables (LR/NB) was performing as well as a model which accounts for interactions or conditional dependencies (RF/BN). Based on these results, one could argue that a simple linear score (LR) with fewer allergen components (from 7 to 18) may be as effective as a more complex model. However, we cannot rule out the possibility that interactions among allergens may have a potentially important role.

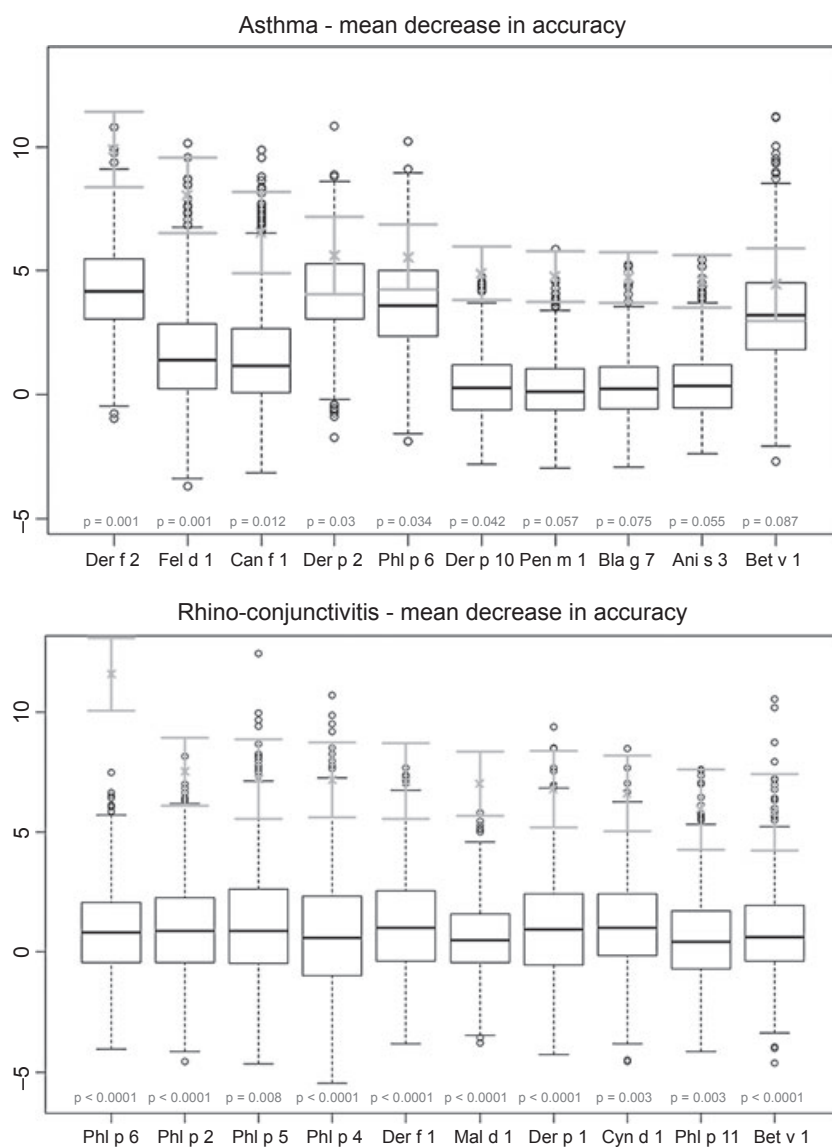


Figure 2 Feature importance plots for the asthma outcome (upper panel) and for the rhino-conjunctivitis outcome (lower panel) measured as mean decrease in accuracy from fitting a random forest and performing an outcome permutation test (1000 runs). Green intervals represent rescaled average ($\pm$ standard deviation) decrease in accuracy, whilst box plots represent the null distribution (randomized outcomes); p-values are highlighted in red. Only the first 10 variables shown.

Our data suggest that sIgE discretization and transformation policy may increase the model performance compared with the single dichotomous threshold at 0.3 ISU. The number of positive sIgEs and the sum of all sIgE levels were poorer predictors of asthma, wheeze and AHR compared with the information on all components, supporting the evaluation of all specific components rather than an overall qualitative assessment.

Limitations and interpretation

From a methodological point of view, one limitation of this study lies within the procedures for feature/model selection. A challenge within BN learning is the simultaneous estimation of both node set and topology, limited here by the usage of two nested heuristic searches. Reliability of the associations in the networks was not assessed, nor the stability of selected feature sets. The selection of main effects and

interactions might be also dependent on the discretization policy.

The advantage of coupling machine learning methods with classical statistical models is that more complex mechanistic hypotheses can be investigated and that large, sparse, heterogeneous data sets can be analysed. Furthermore, by comparing performance of nonlinear and linear methods, one can ascertain whether the unexplained variability (e.g., poor prediction of the outcome) can be reduced by measuring other potentially important variables, permitting formulation of new hypotheses.

Our data suggest that a careful and informed sIgE discretization/transformation may increase the performance, although the differences in AUROC of various approaches could not be always confirmed at the formal 0.05 significance level. The semiquantitative coding suggested by the manufacturer performed better than using the binary threshold at 0.3 ISU (albeit not significantly). In our data set, the automated supervised discretization method yielded the best

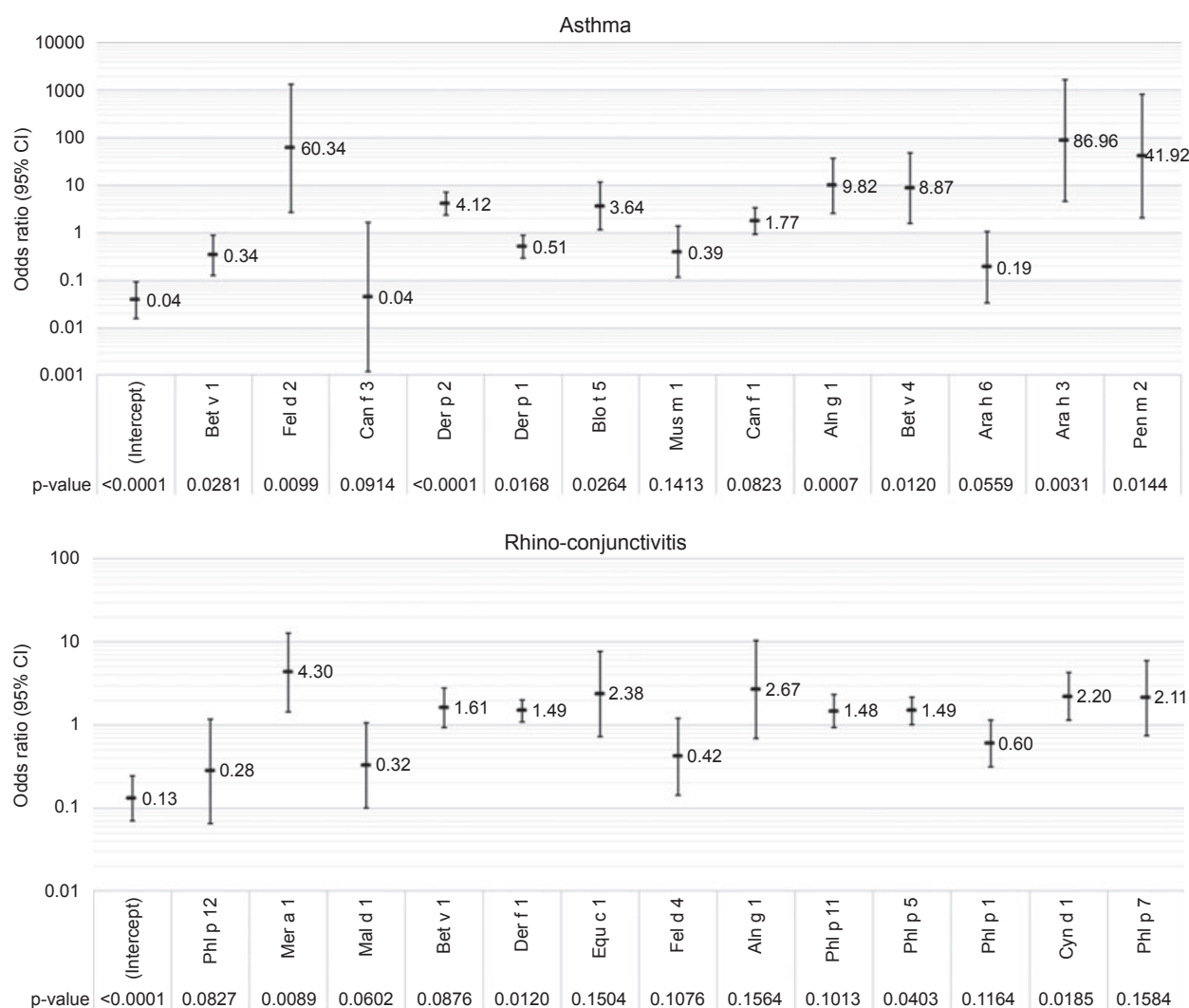


Figure 3 Multivariable logistic regression for asthma and rhino-conjunctivitis outcomes (upper and lower panel, respectively), showing mutually adjusted odds ratios and associated p-values from the LogitBoost algorithm (run on the whole data set). Only variables significant in the univariate analysis were included ($p < 0.05$).

results, supporting the notion that a range of population-specific (and perhaps age-specific) expected values should be established for different populations. Given the multimodal and skewed nature of sIgE distributions and relatively modest sample size, further analyses in larger data sets are warranted. We cannot exclude the possibility that different thresholds are applicable to different components.

The heuristic algorithms for attribute selection yielded compact sets of predictors (from hundreds to a dozen, confirmed when applying the automated supervised discretization approach), which simplifies the interpretation of the results. Not surprisingly, we confirmed previous findings (12) that mite, pollen and pet allergens are top-scoring predictors of asthma, whilst pollens (and mite) are for rhino-conjunctivitis. However, execution of the feature selection algorithms multiple times led to variations in the sets and scores. This may be in part due to highly correlated variables, but also other latent

causes. Such variability does not permit identification of one unique model, and therefore, a unique pathway. Our results suggest that pre-clustering of allergens into relevant families may help stabilize the feature selection phases and may potentially be more useful than standard classifications by source (e.g., pollen or mite).

The prediction models for eczema had poor performance, suggesting either that disruption in skin barrier function may be more important in the pathogenesis of eczema than IgE-mediated mechanisms or that other allergens not available on the chip may be predictive of eczema (e.g., *Staphylococcus aureus* enterotoxins) (23).

Conclusions

Component-resolved diagnostic tests may offer a more accurate assessment of allergic diseases. However, further

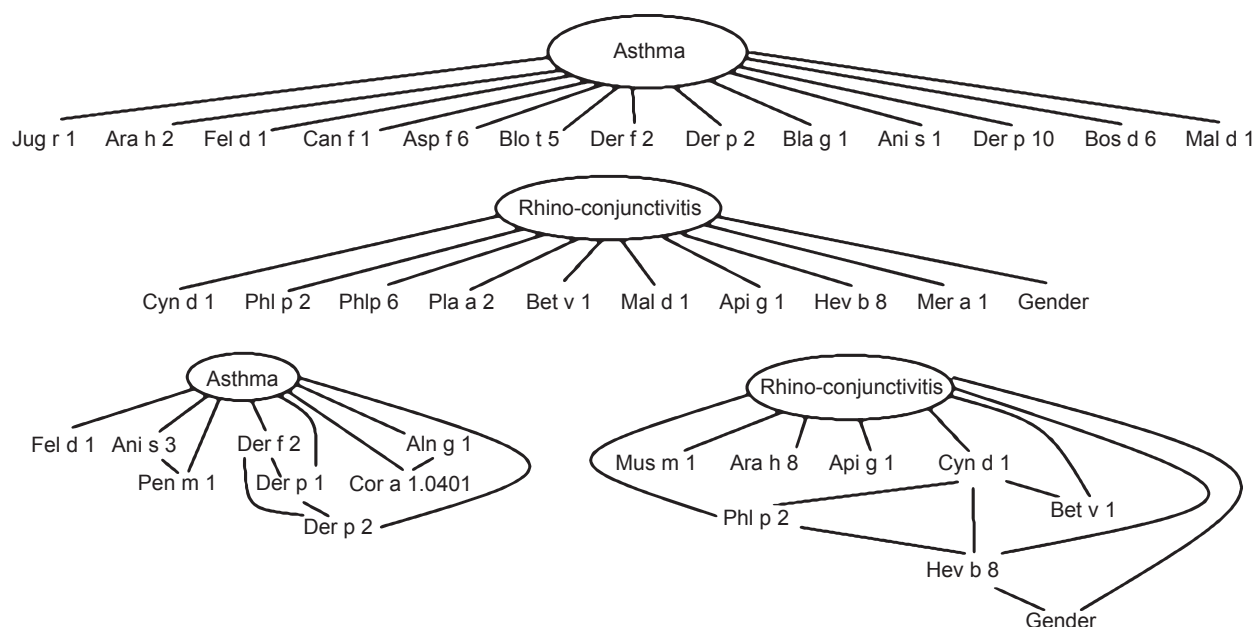


Figure 4 Bayesian networks (BN) for the classification of asthma and rhino-conjunctivitis. Upper panel shows the naïve Bayes (NB) models (hypothesizing variable independence, it can be abstracted to a main-effect logistic model, that is, a linear score where each variable has a weight), and lower panel shows the BN that allow for more complex (direct and indirect) conditional dependencies. Given the non-superiority of the more complex BN model as compared to the NB on the current data set, one could choose the NB hypothesis and further evaluate different variable sets, as in main-effects logistic regression.

improvements in threshold ascertainment and/or value transformations for different allergen components, well-thought-out interpretation algorithms and selection of components are needed to fully capitalize on the potential of the technology.

Authors contributions

MCFP machine learning modelling, manuscript writing; DB data extraction, statistics, statistical review, manuscript review; AS study cohort management, experimental set up, manuscript review; AC principal investigator, study design, manuscript review; IB statistical review, manuscript review.

Conflict of interest

The authors declare no conflict of interest in relation to this study.

References

- Burrows B, Martinez FD, Halonen M, Barbee RA, Cline MG. Association of asthma with serum IgE levels and skin-test reactivity to allergens. *N Engl J Med* 1989; **320**: 271–7.
- Beeh KM, Ksoll M, Buhl R. Elevation of total serum immunoglobulin E is associated with asthma in nonallergic individuals. *Eur Respir J* 2000; **16**: 609–14.
- Simpson BM, Custovic A, Simpson A, et al. NAC Manchester Asthma and Allergy Study (NACMAAS): risk factors for asthma and allergic disorders in adults. *Clin Exp Allergy* 2001; **31**: 391–9.
- Marinho S, Simpson A, Söderström L, Woodcock A, Ahlstedt S, Custovic A. Quantification of atopy and the probability of rhinitis in preschool children: a population-based birth cohort study. *Allergy* 2007; **62**: 1379–86.
- Simpson A, Soderstrom L, Ahlstedt S, Murray CS, Woodcock A, Custovic A. IgE antibody quantification and the probability of wheeze in preschool children. *J Allergy Clin Immunol* 2005; **116**: 744–9.
- Taylor PE, Jacobson KW, House JM, Glovsky MM. Links between pollen, atopy and the asthma epidemic. *Int Arch Allergy Immunol* 2007; **144**: 162–70.
- Gent JF, Belanger K, Triche EW, Bracken MB, Beckett WS, Leaderer BP. Association of pediatric asthma severity with exposure to common household dust allergens. *Environ Res* 2009; **109**: 768–74.

8. Wang J, Calatroni A, Visness CM, Sampson HA. Correlation of specific IgE to shrimp with cockroach and dust mite exposure and sensitization in an inner-city population. *J Allergy Clin Immunol* 2011; **128**: 834–7.
9. Sordillo JE, Webb T, Kwan D, et al. Allergen exposure modifies the relation of sensitization to fraction of exhaled nitric oxide levels in children at risk for allergy and asthma. *J Allergy Clin Immunol* 2011; **127**: 1165–72.e5.
10. De Knop KJ, Bridts CH, Verweij MM, et al. Component-resolved allergy diagnosis by microarray: potential, pitfalls, and prospects. *Adv Clin Chem* 2010; **50**: 87–101.
11. Melioli G, Bonifazi F, Bonini S, et al. The ImmunoCAP ISAC molecular allergology approach in adult multi-sensitized Italian patients with respiratory symptoms. *Clin Biochem* 2011; **44**: 1005–11.
12. Patelis A, Gunnbjörnsdóttir M, Malinovschi A, et al. Population-based study of multiplexed IgE sensitization in relation to asthma, exhaled nitric oxide, and bronchial responsiveness. *J Allergy Clin Immunol* 2012; **130**: 397–402.e2.
13. Custovic A, Simpson BM, Murray CS, Lowe L, Woodcock A. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatr Allergy Immunol* 2002; **13** (Suppl 1): 32–7.
14. Crapo RO, Casaburi R, Coates AL, et al. Guidelines for methacholine and exercise challenge testing-1999. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. *Am J Respir Crit Care Med* 2000; **161**: 1023–9.
15. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajcsy R, ed. *Proceedings of the International Joint Conference on Uncertainty in AI*. San Francisco, USA: Morgan Kaufmann, 1993: 1022–7.
16. Huber W, Von Heydebreck A, Sülthmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002; **18**(Suppl 1): S96–104.
17. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19**: 185–93.
18. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann Stat* 2000; **28**: 337–407.
19. Murthy SK. Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Min Knowl Disc* 1998; **2**: 345–89.
20. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
21. Millán E, Loboda T, Pérez-de-la-Cruz JL. Bayesian networks for student model engineering. *Comput Educ* 2010; **55**: 1663–83.
22. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010; **26**: 1340–7.
23. Semic-Jusufagic A, Bachert C, Gevaert P, et al. Staphylococcus aureus sensitization and allergic disease in early childhood: population-based birth cohort study. *J Allergy Clin Immunol* 2007; **119**: 930–6.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Example of a decision tree, classifying an outcome of interest (1/2) based on the values of variables x , y and z .

Figure S2. Distribution of IgE values from the ISAC chip (ISU scale) by selecting all (left panels) and strictly positive (>0 , right panel) values.

Figure S3. Performance of RF and LR models by varying the input IgE transformation function, specifically (from left to right): binary 0.3 ISU threshold (blue colour), column-wise

quartile discretisation, matrix-wise quartile discretisation, square-root transform, raw values, supervised discretisation, matrix-wise standardisation, quantile normalisation, and manufacturer's semi-quantitative scale.

Figure S4. Correlation plots for allergen IgEs. Spearman rank-correlation was used, and the correlation matrix has been sorted to group highly-correlated allergen groups.

Figure S5. Hierarchical clustering of allergen IgEs. Manhattan ($L1$ norm) distance was used, aggregating instances progressively with the complete linkage method.

Figure S6. Thresholded adjacency graph of allergen IgE, based on Spearman's rank-correlation.

CHALLENGES IN INTERPRETING ALLERGEN MICROARRAYS IN RELATION TO CLINICAL SYMPTOMS: A MACHINE LEARNING APPROACH

SUPPLEMENTARY MATERIAL

METHODS

Study design: Unselected birth cohort study.

Setting: Maternity catchment area of Wythenshawe and Stepping Hill Hospitals, a stable mixed urban-rural population.

Screening & Recruitment: All pregnant women were screened for eligibility at antenatal visits (8th-10th week of pregnancy). Of the 1,499 couples who met the inclusion criteria, 288 declined to take part in the study and 27 were lost to follow-up between recruitment and birth of a child. A total of 1,184 participants had at least some evaluable data.

Clinical Follow-up: Children attended review clinic at age 11 years. A validated ISAAC questionnaire was interviewer-administered.

Airway hyper-reactivity: Airway reactivity was assessed by methacholine challenge using a 5 step protocol performed according to ATS guidelines. Quadrupling doses of methacholine (0.0625–16.0 mg/mL) were delivered to subjects via a DeVilbiss 646 nebuliser (Sunrise Medical HHG, Somerset, PA) and a KoKo dosimeter (Pulmonary Data Services, Doylestown, PA) calibrated to deliver 0.009 mL per 0.6s actuation. The predicted FEV₁ was calculated and if the measured value was <1.0 L or less than 60% predicted the test was not performed. FEV₁ was measured 30 and 90 seconds after 5 inhalations of each dose of methacholine. The challenge was stopped when either a 20% fall in FEV₁ was observed, or the maximum methacholine concentration had been administered.

Statistical learning

Logistic regression (LR)

We fitted main-effects logistic regression (LR) models using (i) either the sum of and the number of positive sIgE values, or (ii) sIgE to all allergen components. For the latter, due to the high number of variables involved, LR was subject to feature selection *via* cross-validated LogitBoost (1). We expressed sIgE values as:

- 1) discretised using the threshold of 0.3 ISU (binary)

- 2) discretised using the manufacturer's semi-quantitative scale as: 0=undetectable or very low, <0.3 ISU; 1=low, ≥ 0.3 and <1 ISU; 2=moderate to high, ≥ 1 and <15 ISU; 3=very high, ≥ 15 ISU).
- 3) discretised using an automated supervised discretisation approach (2)
- 4) continuous raw values
- 5) their square-root/hyperbolic-arcsine transformation (3)
- 6) other microarray normalisation methods such as quantile normalisation (4)

Decision tree (DT) and random forest (RF) models

Decision trees are machine learning methods that divide the population into nested subgroups according to values of the covariates, usually those that have the highest discriminatory power with respect to the outcome of interest (e.g. according to a chi-square test). For instance, the original population of children in our study could be divided into two subgroups according to a Fel d 1 IgE value below or above 0.3 ISU. Then other tree-branching (or "split") rules can be inferred on the two sub-populations, and so on recursively until a stopping criterion (called "pruning") is met (e.g. a minimum number of subjects per subgroup). This progressive data partition can be represented in the form of a tree (Figure S1). Prediction rules are placed at the leaves of this tree (usually a constant value or a class category, but also a linear combination of input covariates). DTs are therefore easy to interpret, but sometimes have poor predictive power. In the current study, we based DT splits on information gain; pruning (removing sections of the tree that provide little power to classify instances) was used to reduce over-fitting.

Random forests are ensemble of several different DTs, fitted with variable parameters, with the aim to improve prediction performance by combining many decision pathways, e.g. averaging across all tree predictions. Specifically, a single DT within a RF is grown on a bootstrap sample of the original population, and splits are calculated in the same way except for the fact that only a random subset of variables is considered at each split, and no pruning is performed. In a RF, predictions are calculated by combining all single predictions as output by its internal DTs. Usually the average/median of all predictions or the majority vote is used. Here, RF was set up with 500 non-pruned random trees and $\sqrt{P}$ features randomly selected at each split. Feature importance for RF was measured as rescaled mean decrease in accuracy (i.e. mean decrease divided by standard error) by considering RF fits randomising a variable of interest. Importance values were compared against a permutation of the outcome labels (1,000 times) in order to ascertain bias coming from different categorical scales or multimodal distributions (5), selecting variables with an average p-value ≤ 0.05 according to the novel method of Altmann *et al.* (6). The advantage of using this method for variable importance evaluation instead of, for example, showing odds ratio from

main-effects LR, resides in the fact that RF can capture complex variable interactions without the need of explicitly defining them. Therefore, the importance of a variable that has an effect only in combination with one other (or more) can be assessed.

Bayesian Networks (BN)

Bayesian Networks were fit here with the aim to describe in an interpretable way (i.e. via graphs) both the variables directly associated with the outcomes and the indirect dependencies across variables and outcomes. BN are directed acyclic graphs in which each node is a covariate and a directed link between two nodes represent a causal relation. If no link is present between two nodes, then they are conditionally independent. Therefore, a variable is uniquely determined by the values of its parents in the network. A “triangulation” or a “cascade” of links can be interpreted as the presence of confounding factors or indirect associations. A BN can be oriented towards an outcome node and inference can be made. With some restrictions (given by unobserved variables and network topology characteristics), a BN explains causal relationships, and even under relaxed hypotheses a BN is able to represent direct/indirect associations among variables. An intuitive and comprehensive introduction of BN modelling for biomedical sciences has been given by Millán et al. (7).

The *naïve Bayes* (NB), a simpler model which assumes conditional independence among variables, was also fit as a control to BN. NB is “conceptually” equivalent to a main-effect LR model. In fact, although NB and LR optimise different loss functions, under certain circumstances the two may converge to the same solution (8). Instead, a BN relaxes the independence assumption; each node variable is independent of its non-descendants in the graph given the state of its parents.

The topology optimisation of BN was achieved using the K2 heuristic algorithm (9), scoring topologies by means of the Bayesian Information Criterion (BIC), and setting the maximum numbers of parents to five. Feature selection for both BN and NB was done by 5-fold cross-validated wrapper method (10), subject to a greedy hill climbing search over the space of attribute subsets augmented with a backtracking facility (11).

Model performance

Goodness-of-fit functions for assessing prediction performance were:

- 1) area under the receiver operating characteristic (AUROC), which is equal to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one;
- 2) sensitivity, i.e. the probability that a test result is positive when the condition is present (true positive rate);

- 3) specificity, i.e. the probability that a test result is negative when the condition is not present (true negative rate) (12).

Sensitivity and specificity were calculated using a fixed output probability threshold of 0.5. The robustness of model performance was analysed *via* repeated validation, and the goodness-of-fit distributions from test splits were compared using the Nadeau and Bengio's t-test, which accounts for a modification of the degrees of freedom given the entity of sample overlap (13,14).

Sample size

Analytic formulation for determining the sample size for DT and RF is not available (15). For LR, we used the formula of Peduzzi *et al.* (16) where the sample size $N=10*k/p$ (k being the number of covariates and p the smallest proportion of negative/positive cases in the outcome variable); therefore, with 238 cases, up to 5 covariates can be put in a LR model (simulations made at the 90% power reference).

Variable correlation and cluster analysis

In order to study the effect of correlated variables in the models, we employed unsupervised learning techniques, specifically: hierarchical clustering, rank-correlation analysis (17), and network analysis (18).

RESULTS

Distribution of sIgE values

The distribution of IgE values (after excluding columns with constant zeros) was highly skewed: by dividing it into zero and non-zero values (24,276 vs. 2,478), the latter appeared symmetrical if log-transformed, yet failing the Shapiro-Wilk test for normality, with a median (interquartile, IQR) range of 2.5 (0.65-10.5) ISU. Figure S2 shows the histograms of zero and non-zero values, upon several input transformations.

Transformation of sIgE values

By a preliminary test on variable encodings subject to model performance (AUROC with respect to asthma, using LR and RF), we found that the binary discretisation of IgE values using the 0.3 ISU threshold was performing less well than using a continuous scale or a multiple categorisation (e.g. automated supervised discretisation approach, square-root transform, normalisation of the raw IgE values, et cetera). The automated supervised discretisation method yielded the best results and categorised 17/112 variables with a non-constant indicator, excluding the others, thus this discretisation worked as a preliminary feature selection. However, the decrease in AUROC average of all input transformations as compared to that of the supervised discretisation approach could not be considered significant at the 0.05 level (p-value=0.1, comparing binary 0.3 threshold vs. supervised discretisation). The manufacturer's semi-quantitative scale was better than the binary threshold as well, although not significantly.

Statistical learning models for association with clinical outcomes

Figure 4 in the manuscript depicts two optimised naïve Bayes and Bayesian network structures for asthma and rhino-conjunctivitis. The NB model hypothesises variable independence, and as such can be abstracted to a main-effect logistic model (i.e. a linear score where each variable has a weight). The BN allows more complex (direct and indirect) conditional dependencies. Given the non-superiority of the more complex BN model as compared to the NB on the current data set, one could choose the NB hypothesis and further evaluate different variable sets.

Variable correlation and cluster analysis

Since all the modelling techniques yielded non uniform feature sets over different validation runs, we sought to ascertain if this was partly due to highly-correlated variables. By applying rank-correlation analysis (Figure S4), hierarchical clustering (Figure S5) and network analysis (Figure S6), we identified several clusters of variables with a high-mutual

association. The association was confirmed by using different metrics, including the Spearman's rank-correlation and the Manhattan distance. The most stable groups were:

- (a) Cyn d 1, Phl p 1, Phl p 2, Phl p 4, Phl p 5, Phl p 6
- (b) Der f 1, Der f 2, Der p 1, Der p 2
- (c) Bet v 2, Hev b 8, Mer a 1, Phl p 12
- (d) Bet v 4, Phl p 7
- (e) Ani s 3, Bla g 7, Der p 10, Pen m 1
- (f) Ara h 1, Ara h 2, Ara h 6
- (g) Aln g 1, Bet v 1, Cor a 1.0101, Cor a 1.0401, Mal d 1, Pru p 1

These groups partly explain the variability in the feature sets selected by LR or BN/NB, and the large range of importance values yielded by the RF. For instance, referring to allergens of the group (g), Cor a 1.0401 was the variable with the highest association ($p=0.00003$) with the asthma outcome. Indeed, Cor a 1.0401 was not always selected by LogitBoost LR (only 20% of times across all validation runs), but alternatively Bet v1 and Aln g1 could be picked up. Likewise, the RF ranked Bet v 1 among the top-scoring variables, whilst Mal d 1 was selected by the NB model. In fact, the BN detected an association of both Aln g 1 and Cor a 1.0401 with the outcome, and also an association between the two allergens.

Figure S1. Example of a decision tree, classifying an outcome of interest (1/2) based on the values of variables x, y and z. At each tree node the population is divide into subgroups according to a test on the variable split. The leaf nodes represent classification rules calculated from the root to the tip of the tree. For instance the rightmost pathway represent the rule “IF (condition c for variable x is met) AND IF (condition g for variable z is met) THEN (probability of outcome 1 is 97%)”. Therefore the tree is able also to code variable interactions.

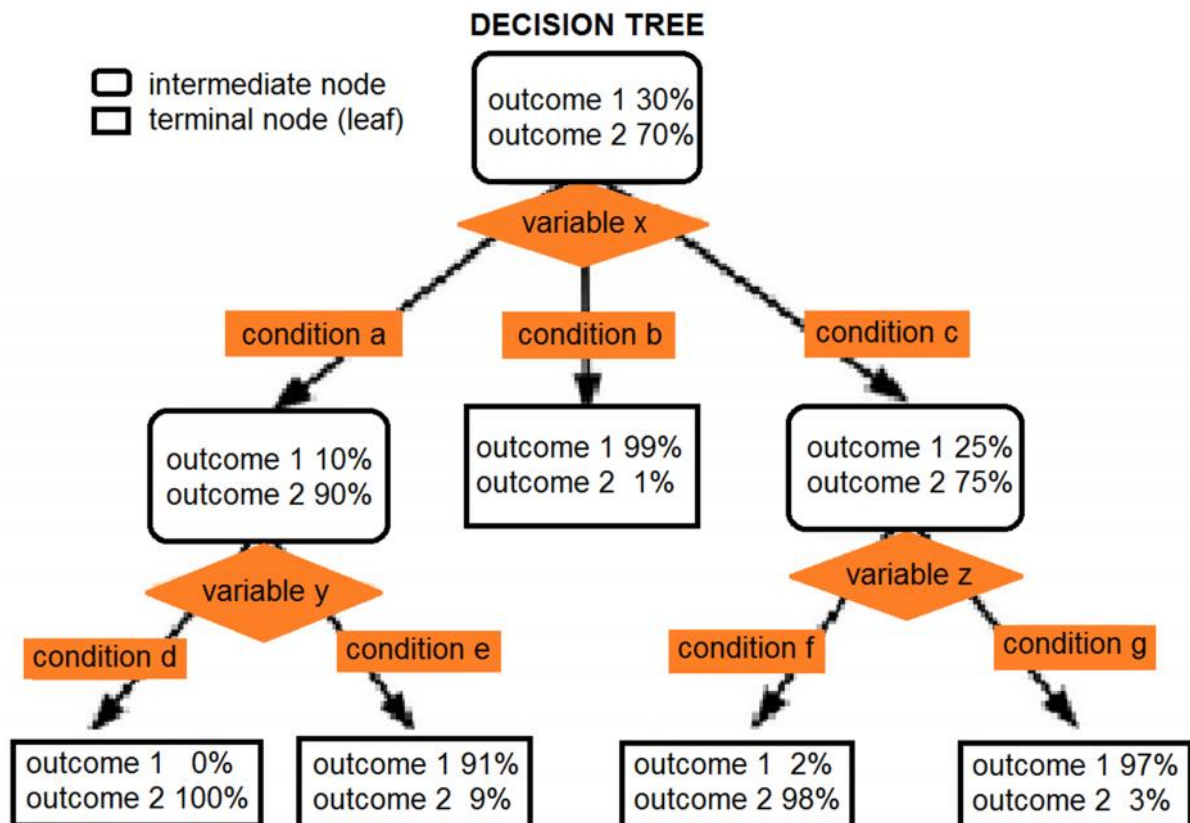


Figure S2. Distribution of IgE values from the ISAC chip (ISU scale) by selecting all (left panels) and strictly positive (>0 , right panel) values. Different non-linear transformations are shown.

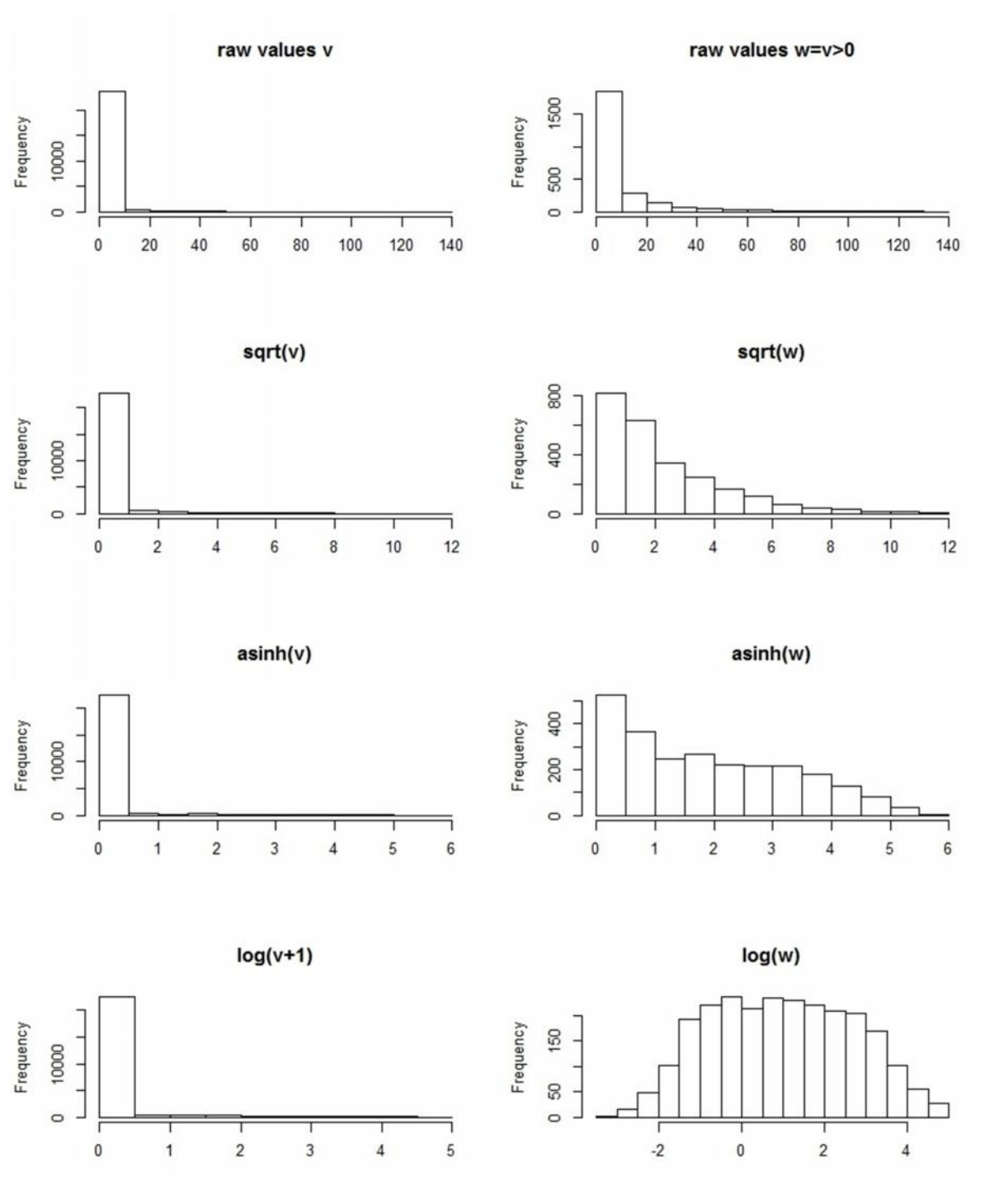


Figure S3. Performance of RF and LR models by varying the input IgE transformation function, specifically (from left to right): binary 0.3 ISU threshold (blue colour), column-wise quartile discretisation, matrix-wise quartile discretisation, square-root transform, raw values, supervised discretisation, matrix-wise standardisation, quantile normalisation, and manufacturer's semi-quantitative scale. Boxplots represent out-of-bag AUROCs averaged across 50 bootstrap runs, and p-values are calculated with respect to the RF trained using the supervised discretisation (the best model in terms of AUROC). Green boxplots are from RF and LR trained on the supervised discretisation, whilst blue boxplots represent AUROCs from the binary discretisation at 0.3 ISU (p-value=0.1). Red boxplots refer to the manufacturer's semi-quantitative scale.

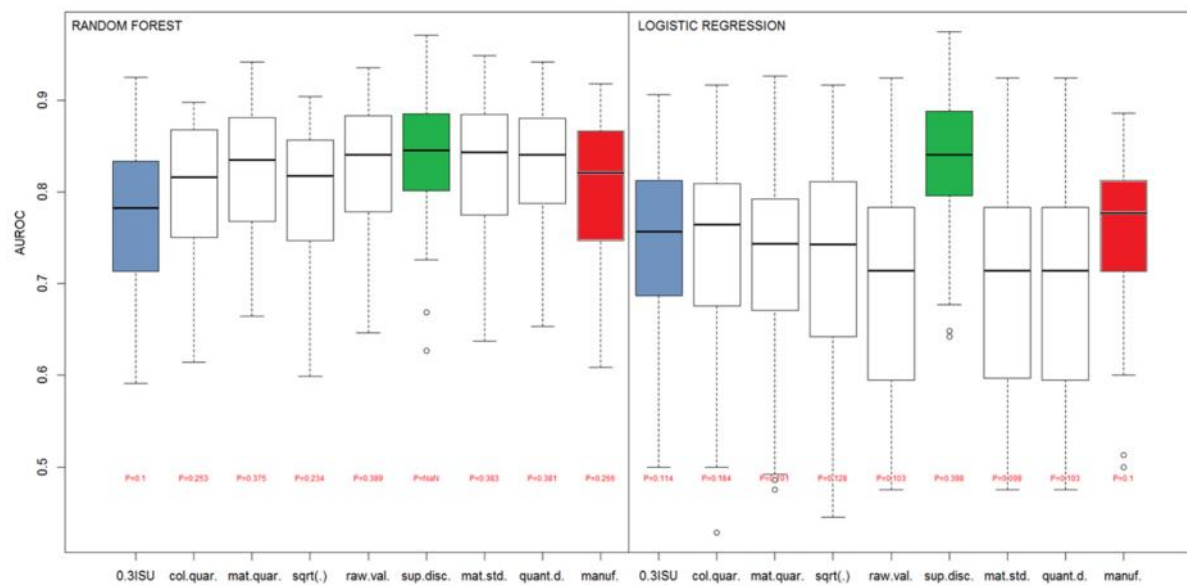


Figure S4. Correlation plots for allergen IgEs. Spearman rank-correlation was used, and the correlation matrix has been sorted to group highly-correlated allergen groups. Deep blue represents high-level correlation, deep red is the opposite, whilst white means no correlation.

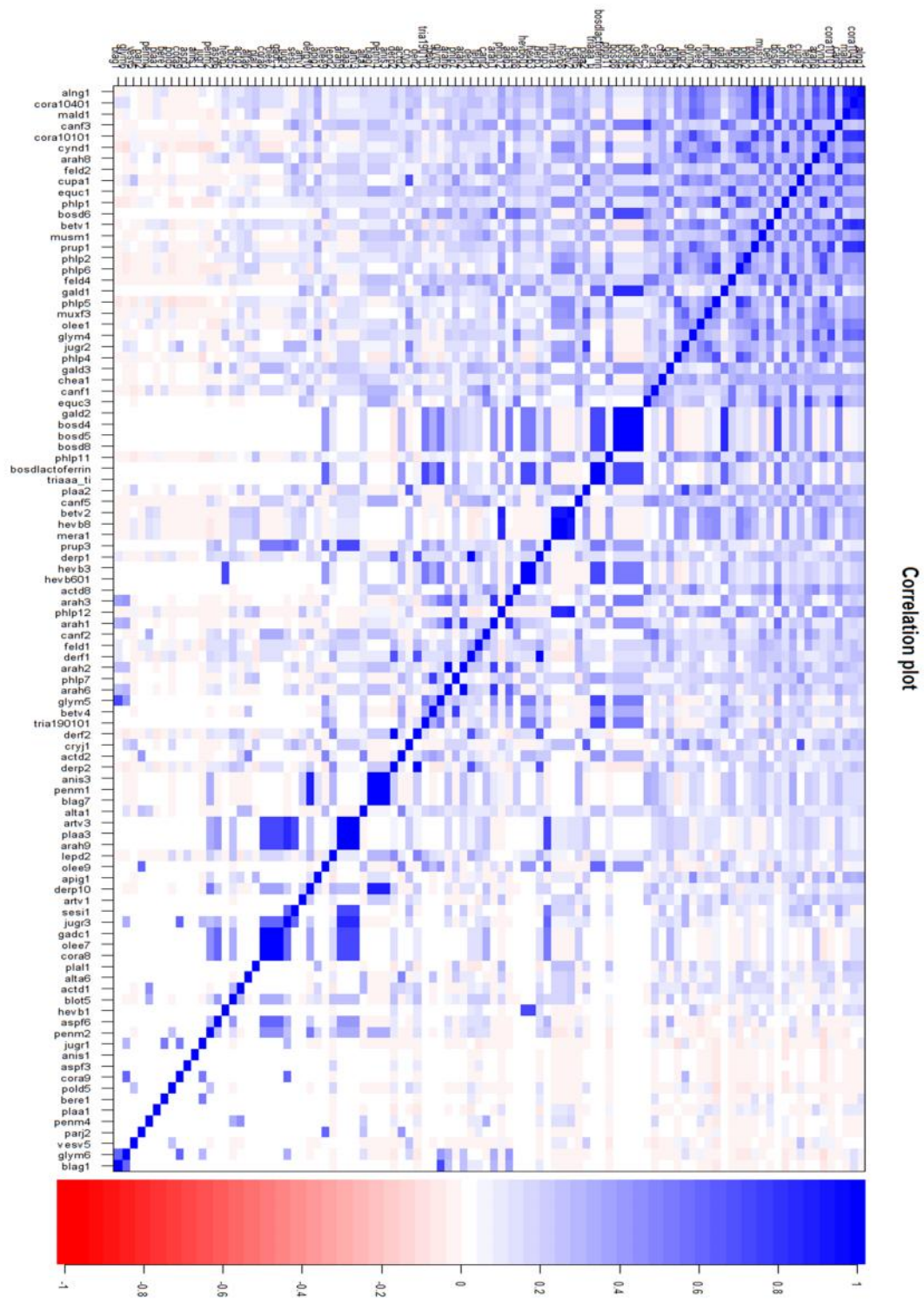


Figure S5. Hierarchical clustering of allergen IgEs. Manhattan (L1 norm) distance was used, aggregating instances progressively with the complete linkage method. Bootstrap analysis was performed (500 resampled sets), and node reliability is indicated in % as red/green p-values/bootstrap support.

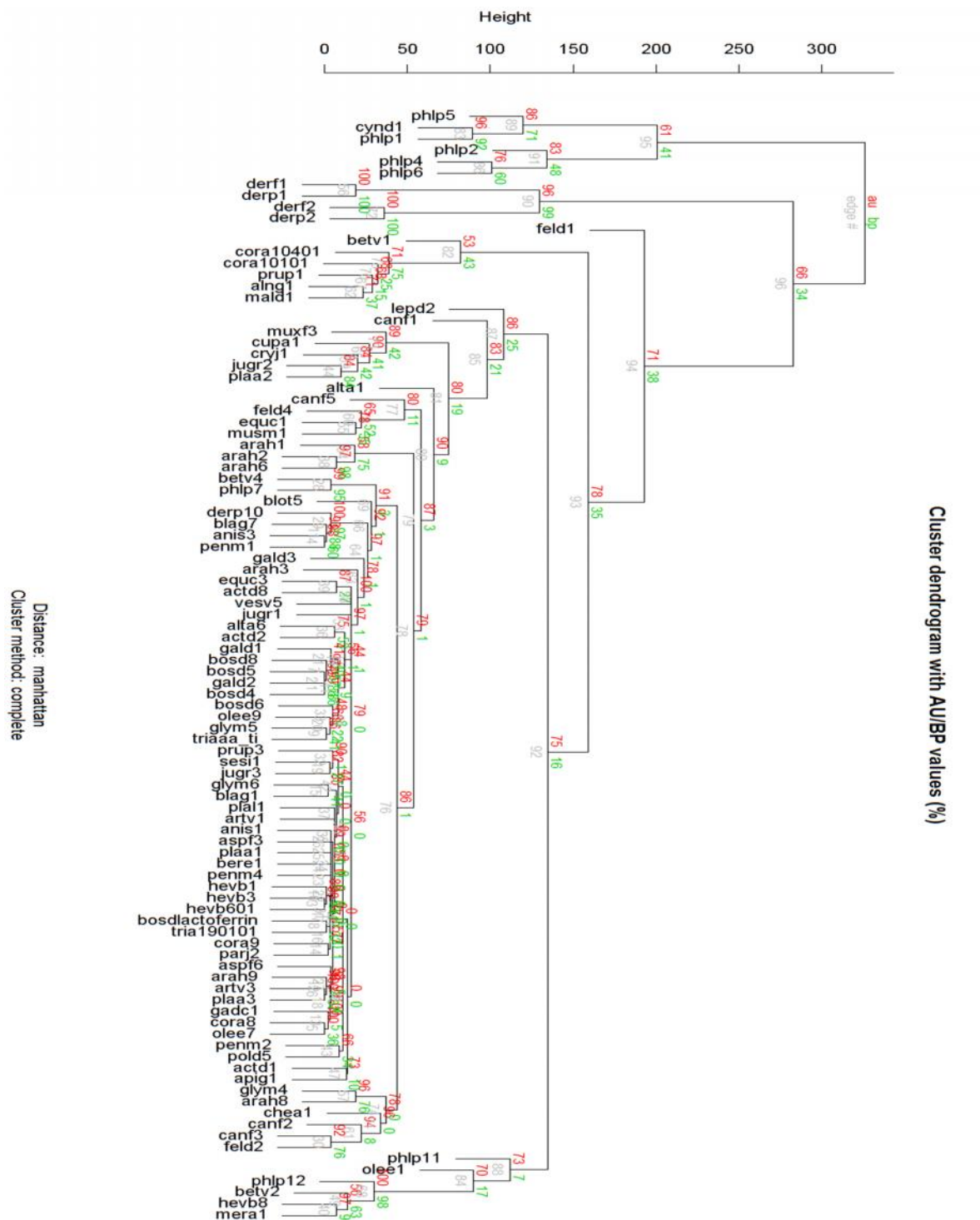


Figure S6. Thresholded adjacency graph of allergen IgE, based on Spearman's rank-correlation. Variables with a correlation >95th percentile of the whole set of correlations across all data are connected. Singletons represent variables that have no relevant correlation with the others.

REFERENCES

1. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics* 2000;**28**:337–407.
2. Fayyad UM, Irani KB. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In: Bajcsy R, ed. *Proceedings of the International Joint Conference on Uncertainty in AI*. Morgan Kaufmann 1993. 1022–7.
3. Huber W, Von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002;**18 Suppl 1**:S96–S104.
4. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**:185–93.
5. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Strobl, Boulesteix, Zeileis, Hothorn: Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *Online* 2006;**490**:1–6.
6. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010;**26**:1340–7.
7. Millán E, Loboda T, Pérez-de-la-Cruz JL. Bayesian networks for student model engineering. *Computers & Education* 2010;**55**:1663–83.
8. Ng AY, Jordan MI. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems* 2002;**2**:841–8.
9. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 1992;**9**:309–47.
10. Kohavi R, John GH. Wrappers for feature subset selection. *Artificial Intelligence* 1997;**97**:273–324.
11. Russell SJ, Norvig P. Artificial Intelligence, A Modern Approach - 2nd Edition.pdf. *Artificial Intelligence*. 2003;**82**:874.
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. *Elements* 2009;**1**:337–87.
13. Bengio Y, Grandvalet Y. Bias in Estimating the Variance of K-Fold Cross-Validation. In: Duchesne P, Remillard B, eds. *Statistical Modeling and Analysis for Complex Data Problem*. Springer 2005. 75–95.
14. Nadeau C, Bengio Y. Inference for the Generalization Error. *Machine Learning* 2003;**52**:239–81.
15. Sug HSH. A Structural Sampling Technique for Better Decision Trees. *Ieee* 2009.

16. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996;**49**:1373–9.
17. Hahne F, Huber W, Gentleman R, Falcon S, Gentleman R, Carey VJ. Unsupervised Machine Learning. *Bioconductor Case Studies*. 2008;:137–57.
18. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Systems* 2006;**Complex Sy**:1695.