

Accepted Manuscript

Novel Sensitization Trajectories in Childhood revealed by a Cluster Analysis

Ann-Marie M. Schoos, MD, PhD, Bo L. Chawes, MD, PhD, DMSc, Erik Melén, MD, PhD, Anna Bergström, PhD, Inger Kull, PhD, Magnus Wickman, MD, PhD, Klaus Bønnelykke, MD, PhD, Hans Bisgaard, MD, DMSc, Morten A. Rasmussen, PhD



PII: S0091-6749(17)30475-X

DOI: [10.1016/j.jaci.2017.01.041](https://doi.org/10.1016/j.jaci.2017.01.041)

Reference: YMAI 12708

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

Received Date: 1 May 2016

Revised Date: 19 December 2016

Accepted Date: 4 January 2017

Please cite this article as: Schoos A-MM, Chawes BL, Melén E, Bergström A, Kull I, Wickman M, Bønnelykke K, Bisgaard H, Rasmussen MA, Novel Sensitization Trajectories in Childhood revealed by a Cluster Analysis, *Journal of Allergy and Clinical Immunology* (2017), doi: 10.1016/j.jaci.2017.01.041.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. 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.

Novel Sensitization Trajectories in Childhood revealed by a Cluster Analysis

Ann-Marie M Schoos¹, MD, PhD; Bo L Chawes¹, MD, PhD, DMSc; Erik Melén^{2,3}, MD, PhD;
Anna Bergström², PhD; Inger Kull^{3,4}, PhD; Magnus Wickman^{2,3}, MD, PhD; Klaus Bønnelykke¹,
MD, PhD; Hans Bisgaard¹, MD, DMSc; Morten A Rasmussen^{1,5}, PhD

Affiliations:

¹ Copenhagen Prospective Studies on Asthma in Childhood, Faculty of Health and Medical
Sciences, University of Copenhagen & Danish Pediatric Asthma Center, Gentofte Hospital,
University of Copenhagen; Denmark.

² Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

³ Sachs Children's Hospital, Södersjukhuset, Stockholm, Sweden.

⁴ Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet,
Stockholm, Sweden

⁵ Department of Food Science, Faculty of Science, University of Copenhagen; Denmark.

Correspondence:

Professor Hans Bisgaard, MD, DMSc

Copenhagen Prospective Studies on Asthma in Childhood, Danish Pediatric Asthma Center,
Gentofte Hospital, University of Copenhagen

Ledreborg Alle 34

DK-2820 Gentofte; Denmark

Tel: (+45) 39777360

Fax: (+45) 39777129

E-mail: bisgaard@copsac.com

Website: www.copsac.com

Contributions: The guarantor of this study is HB who is responsible for the integrity of the work as a whole, from conception and design to conduct of the study and acquisition of data, analysis and interpretation of data and writing of the manuscript. AMS and MR are responsible for data acquisition, analysis, interpretation and writing the manuscript. EM, AB, IK and MW are responsible for the acquisition of data and interpretation of the results generated from BAMSE. All co-authors have contributed substantially to the analysis and interpretation of the data, and have provided important intellectual input and approval of the final version of the manuscript.

Source of funding: COPSAC is funded by private and public research funds all listed on www.copsac.com. The Lundbeck Foundation; Danish State Budget; Danish Council for Strategic Research; Danish Council for Independent Research and The Capital Region Research Foundation have provided core support for COPSAC. BAMSE is funded by The Swedish Research Council, The Swedish Heart-Lung Foundation, Stockholm County Council (ALF) and the Swedish Research Council Formas. The funding agencies did not have any role in design and conduct of the study; collection, management, and interpretation of the data; or preparation, review, or approval of the manuscript. No pharmaceutical company was involved in the study.

Running Head: Sensitization patterns in childhood

Article type: Original article

Word count: 3,799

46 Word count, abstract: 277

47 Figure count: 4

48 Tables: 1

49 Online Repository: Yes

50

51

Abstract

Background

Assessment of sensitization at an isolated time-point during childhood provides limited clinical information. We hypothesized that sensitization develops as specific patterns with respect to age at debut, development over time, and involved allergens, and that such patterns might be more biologically and clinically relevant.

Objective

To explore latent patterns of sensitization during the first 6 years of life and investigate whether such patterns associate to development of asthma, rhinitis, and eczema.

Methods

We investigated 398 children from the at-risk Copenhagen Prospective Studies on Asthma in Childhood 2000 birth cohort (COPSAC₂₀₀₀) with specific-IgE against 13 common food and inhalant allergens at ages ½, 1½, 4, and 6yrs. An unsupervised cluster analysis for three-dimensional data (NNS-PARAFAC) was used to extract latent patterns explicitly characterizing temporal development of sensitization, while clustering allergens and children. Subsequently, these patterns were investigated in relation to asthma, rhinitis, and eczema. Verification was sought in an independent unselected birth cohort, BAMSE, constituting 3051 children with specific-IgE against the same allergens at 4 and 8yrs.

Results

The NNS-PARAFAC analysis indicated a complex latent structure involving seven age- and allergen-specific patterns in the COPSAC₂₀₀₀ data: (1) dog/cat/horse; (2) timothy grass/birch; (3) molds; (4) house dust mites; (5) peanut/wheat flour/mugwort; (6) peanut/soybean; and (7)

egg/milk/wheat flour. Asthma was solely associated with pattern 1 (OR=3.3 [1.5-7.2]), rhinitis with patterns 1-4 and 6 (ORs=2.2–4.3) and eczema with patterns 1-3 and 5-7 (ORs=1.6–2.5). All seven patterns were verified in the independent BAMSE cohort ($R^2>0.89$).

Conclusion

This study suggests the presence of specific sensitization patterns in early childhood differentially associated to development of clinical outcomes. Using such patterns in future research might provide more robust and clinically relevant results.

Keywords: Sensitization, Patterns, Asthma, Rhinitis, Eczema, Specific-IgE, Children

Key Messages

- Sensitization is not a stable dichotomous trait but can be described as 7 different phenotypes defined by the allergen and temporal pattern of sensitization.
- Association with asthma, rhinitis, and eczema depends on the sensitization pattern.

Capsule Summary

Sensitization in young children is a make-up of seven distinct patterns defined by the allergens and temporal profile and with different association to allergic disease.

Abbreviations

BAMSE = Children Allergy Milieu Stockholm Epidemiology [in Swedish]

COPSAC = Copenhagen Prospective Studies on Asthma in Childhood

HR = Hazard Ratio

OR = Odds Ratio

PARAFAC = PARAllel FACtor analysis

NNS-PARAFAC = NonNegative Sparse PARAFAC

sIgE = specific Immunoglobulin E

Introduction

We recently demonstrated an age-related association between allergic sensitizations for both inhalant and food allergens and clinical manifestations of asthma and eczema¹. We therefore hypothesized that age at debut, temporal development, and type and number of allergen sensitizations involved may reflect underlying patterns of allergic sensitization that are closer associated to disease and therefore more clinically relevant.

Recent studies have also suggested the presence of several allergic sensitization patterns in childhood²⁻⁴, however, the previous published methods for time course pattern recognition in the development of sensitization are based on probabilistic hidden Markow models. We wanted to use a statistical model that estimates a low dimensional representation of the observed data, with minimal distributional assumptions. Furthermore, a method that entails the ability to model higher ordered data structures without the need to unfold data into a matrix, and thereby abide to the nature of the data structure.

The aim of this study was to explore latent patterns of sensitization during the first 6 years of life using an unsupervised data driven cluster analysis (NonNegative Sparse PARAllel FACtor analysis – NNS-PARAFAC). We analyzed measurements of specific-IgE (sIgE) against 13 common inhalant and food allergens assessed longitudinally at ages ½, 1½, 4, and 6 years in children participating in the Copenhagen Prospective Study on Asthma in Childhood 2000 (COPSAC₂₀₀₀) birth cohort. We used the NNS-PARAFAC to extract a set of patterns in the complex multidimensional data pool that characterizes longitudinal sensitization patterns in children, unbiased from pre-specified clinical hypotheses. We then related those patterns to development of asthma, rhinitis, and eczema. This method explicitly models the temporal

126 development of sensitization while clustering the allergens and children, and as such proposes a
127 general framework for analysis of longitudinal multivariate data structures.
128 Verification was sought in another independent Scandinavian birth cohort, BAMSE⁵, where the
129 same sIgE assessments were performed at ages 4 and 8 years.
130

Methods

Study Design

COPSAC₂₀₀₀ is a single-center, prospective clinical birth cohort of children born to mothers with a history of asthma enrolled during 1998-2001⁶⁻⁸. The children were closely monitored from birth till age 7 with visits to the research clinic every 6 months and additional visits at the onset of airway or skin symptoms. Asthma, rhinitis, and eczema were diagnosed and monitored at these visits by the COPSAC pediatricians according to standard operating procedures. Diagnoses were defined by exact start and finish dates scrutinized from daily diary cards filled by the parents.

The Copenhagen Ethics Committee (KF 01-289/96) and The Danish Data Protection Agency (2008-41-2434) approved the study, and informed consent was obtained from both parents at enrolment.

BAMSE is a longitudinal population-based birth cohort including 4089 children born between 1994 and 1996 in different areas of Stockholm^{5,9}. At the 4- and 8-year-follow-up, blood samples were taken for sIgE analyses. In this study, we included 3051 children, who provided IgE data at 4 and/or 8 years.

Allergic Sensitization

Blood was sampled at ages ½, 1½, 4, and 6 years (COPSAC₂₀₀₀) and 4 and 8 years (BAMSE) for measurements of sIgE levels (ImmunoCAP, Phadiatop Infant™, Pharmacia Diagnostics AB, Uppsala, Sweden) against 8 inhaled allergens (birch, timothy grass, mugwort, dog, cat, horse, molds, house dust mite (*D. pteronyssinus*)) and 5 food allergens (egg, milk, peanut, soybean, wheat flour). Sensitization was defined as sIgE $\geq$ 0.35 kU_A/L^{6,10}.

Clinical Outcomes

Asthma in COPSAC₂₀₀₀ was diagnosed at age 7 years based on an algorithm including 4 mandatory criteria: (1) five episodes of diary verified troublesome lung symptoms each lasting minimum 3 days within a period of 6 months; (2) symptoms judged by the COPSAC pediatricians to be typical of asthma, e.g. exercise induced symptoms, prolonged nocturnal cough, recurrent cough outside common cold, symptoms causing wakening at night; (3) need of intermittent rescue use of inhaled β 2-agonist; and (4) response to a 3-month course of daily inhaled corticosteroids and relapse after ended treatment⁷.

Rhinitis in COPSAC₂₀₀₀ was diagnosed at age 7 years based on a parental interview on history of symptoms performed at the COPSAC research unit. Rhinitis was defined as significant and reoccurring sneezing, blocked, or runny nose severely affecting the wellbeing of the child in the past 12 months in periods without accompanying common cold or flu, and requiring congruence between symptoms and relevant exposure¹¹⁻¹⁴.

Eczema in COPSAC₂₀₀₀ was diagnosed at scheduled and acute visits at age 0-7 years according to the Hanifin and Rajka criteria⁸ capturing both age of debut and remission as previously detailed⁹. In this study we used age of debut as outcome definition for eczema.

For definition of the BAMSE outcomes see the Online Repository.

Statistical Methods

An unsupervised data driven multidimensional factor analysis technique called the NonNegative Sparse PARAFAC (NNS-PARAFAC) model¹⁵ was applied to analyze the COPSAC₂₀₀₀ data reflecting subjects, age, and type of allergens¹⁶. The method does not entail exclusive

partitioning of the children, but allows a child to belong to several patterns. Similarly a single allergen can contribute to several patterns.

The number of patterns was estimated based on: (1) Scree plot of variance explained; and (2) pattern interpretability, where biological plausible partitioning of the allergens and distinct support of only a few allergens per pattern is sought optimal. Optimization of the sparsity constraint was done by Bayes Information Criterion¹⁷. The method handles missing data via exclusion of these from the objective. Thus, children with missing sIgE assessments were retained in the model if they had at least one sIgE assessment done to conserve maximum study power. For further details please consult the Online Repository.

Primarily, the sensitization patterns were extracted from the COPSAC₂₀₀₀ data and subsequently investigated in relation to asthma and rhinitis by logistic regression and to eczema by Cox regression, both by univariate one-by-one associations, as well as a multivariate model regressing the outcome on all the sensitization patterns, eliminating non significant ($p > 0.05$) patterns by backward elimination. Prior to association analysis each pattern score were dichotomized by setting all non-zero values to one, while leaving the zeros as zeros.

As a post hoc analysis, the estimated NNS-PARAFAC model was used to cluster the children at each time point. This time-dependent clustering was referenced against the clinical outcomes univariately to investigate the age of penetration for each sensitization pattern.

Verification in the BAMSE cohort was conducted by estimation of a NNS-PARAFAC model with the exact same model settings as for the COPSAC model, followed by comparison of the allergen and time patterns and relative size of the individual patterns. Verification of the sensitization pattern versus clinical outcome was conducted by logistic regression for asthma, rhinitis, and eczema at both 4- and 8 years of age, both univariately and multivariately as

described above. These comparisons were primarily conducted on the subset of children of asthmatic mothers in order to mirror the COPSAC₂₀₀₀ study design, and secondly on the entire BAMSE cohort. The robustness of the patterns was investigated towards changes in population due to maternal predisposition, missing sensitization data, and random changes. For details see the Online Repository.

All statistical analyses were conducted in Matlab® R2014b using an *in-house* routine for clustering and statistical toolbox v 8.1 for association analysis between patterns and clinical outcomes. The algorithm is freely available at www.models.life.ku.dk/algorithms.

Results

Baseline Characteristics

In the COPSAC₂₀₀₀ cohort, 398 of the 411 children had at least one available sIgE assessment at one of the four age-points. At age ½yr, 361 (91%) children completed sIgE measurements to all 13 allergens, 355 (89%) at age 1½yrs, 319 (80%) at 4yrs, and 296 (74%) completed at age 6 years. Of the 17,420 sIgE registrations (398 participants x 4 time points x 13 allergens minus 3276 missing entries), 566 were positive (3.2%). For details on the sIgE distributions see **Figure E1**. The distribution of positive tests across the 13 allergens at the four age-points in the COPSAC₂₀₀₀ cohort is shown in **Figure 1**.

In the BAMSE cohort, 3045 of the 4089 children had at least one available sIgE assessment at one of the two age-points. At age 4yrs, 2604 (86%) completed sIgE measurements to all 13 allergens and 2451 (80%) completed at age 8yrs. Of the 65,434 sIgE registrations, 4175 were positive (6.4%).

Table 1 shows the baseline characteristics of children with available sIgE assessments in the two cohorts.

Sensitization Pattern Analysis

The NNS-PARAFAC analysis in the COPSAC₂₀₀₀ birth cohort revealed a total of seven distinct latent patterns: (1) cat/dog/horse (40 children follow this pattern, which explains 16% of the variation in the data); (2) timothy grass/birch (N=57, 10% of the variation); (3) molds (N=12, 7% of the variation); (4) house dust mites (N=39, 7% of the variation); (5) peanut/wheat flour/mugwort (N=45, 11% of the variation); (6) peanut/soybean (N=61, 10% of the variation); and (7) egg/milk/wheat flour (N=108, 8% of the variation). We labeled the patterns based on the

nature of the allergens as *animal* (1), *other airborne* (2-3), *mixed* (5), and *food* (6-7). The patterns are unique and characterized by groups of children, who follow the same sensitization pattern with respect to age at debut, development over time, as well as type of allergen specific sensitizations. These characteristics are outlined in **Figure 2**, which shows both the allergen loadings and the temporal development of the patterns. As an example, **Figure 2A** illustrates that the allergens with the largest influence (highest loadings) in pattern 1 are cat, dog, and horse, which exhibit the same overall profile with respect to participants and time. The time-effect of each pattern is shown in **Figure 2B**, where e.g. sensitivity towards the allergens in pattern 1 is first registered at 1½yrs showing an increase at 4yrs followed by a slight increase at 6yrs.

Verification in the BAMSE Cohort

Figure 2 also shows the comparison of the seven patterns in BAMSE and COPSAC₂₀₀₀, which were in high agreement both in terms of allergen profiles ($R^2 > 0.89$), development over time, and clustering of children, however, with some ambiguity for pattern 3 (*Molds*) (see Online Repository **Table E1**). In general, the patterns are robust towards population changes due to maternal predisposition. However, stratification leads to lack of data support for especially low abundant allergens, resulting in uncertain patterns for these allergens. For details on robustness of the pattern analysis see the Online Repository **Figure E2-7**.

Sensitization Patterns in relation to Clinical Outcomes

The seven longitudinal sensitization patterns were compared with development of asthma, rhinitis, and eczema by pairwise comparisons in COPSAC₂₀₀₀ (**Figure 3**). Asthma was only significantly associated with pattern 1 (cat/dog/horse): OR=3.3; 95% CI, 1.5-7.2. Further, a trend

of association was observed between asthma and pattern 4 (house dust mites): OR=1.9; 95% CI, 0.9-4.4, but pattern 1 remained the primary driver based on the multivariate model (**Table E2**). Rhinitis was significantly associated with patterns 1-4 (cat/dog/horse; timothy grass/birch; molds; house dust mites) and 6 (peanut/soy bean) (ORs=2.2-4.3), of which pattern 2 (timothy grass/birch) and pattern 4 (house dust mites) were found as the primary drivers based on the multivariate model (see **Table E3**). Eczema was found associated with patterns 1-3 (cat/dog/horse; timothy grass/birch; molds) and 5-7 (peanut/wheat flour/mugwort; peanut/soy bean; egg/milk/wheat flour) (HRs=1.6-2.5), from which pattern 1-2 (cat/dog/horse; timothy grass/birch) were the strongest drivers (**Figure 3, Table E4**).

The association analysis was additionally run using a sensitization free control group of n = 249 children without any positive sensitization registrations across all four time points. The results are numerically slightly stronger, due to exclusion of sensitized children from the control group (results not shown).

Results based on the entire BAMSE cohort (n = 3051) showed that all patterns were positively associated with all outcomes at both age of 4 and 8 years (**Figure E8 in the Online Repository**). The associations between asthma and rhinitis and sensitization patterns are strongest at 8 years of age, whereas for eczema the association is strongest at 4 years of age. For the clinical outcome eczema there is a fair agreement between the results obtained in COPSAC₂₀₀₀ and BAMSE, whereas for both asthma and rhinitis this is not the case. Stratifying the BAMSE analysis to only include children of asthmatic mothers (n = 282), and thereby imitating the COPSAC2000 design, changes the results towards higher agreement for asthma and eczema, while rhinitis remains unverifiable. For asthma, five out of seven patterns obtain association results, where the estimate based on COPSAC₂₀₀₀ data at 6yrs is between the BAMSE estimates at 4yrs and 8yrs (**Figure 3**).

Of the three outcomes, only asthma and rhinitis are associated ($p=0.01$) and hence possibly confounds one another in the association analysis. When adjusting the association with rhinitis for asthma the results are unaffected, whereas the pattern-wise association with asthma attenuates when adjusting for rhinitis (see **Table E5** – based on COPSAC₂₀₀₀).

Age of penetration

Figure E9 shows the age dependent, and across *all time* points, cluster sizes for each pattern. In accordance with the raw data, there are only very few children exhibiting sensitization in the first five (non food related) patterns at age 6 months (and to some extent at age 18 months). The results from associating these clusters with the clinical outcomes are shown in **Figure 4** where especially pattern 1 (cat/dog/horse) shows a profile of early sensitization towards these allergens being stronger associated with the clinical outcomes, with a limited amount of cases. As an example, asthma and early life (age 1½y) sensitization promoted the strongest association with an OR of 12.3 [2.2-69.8]. For pattern 6 (peanut/soybean), the opposite relation is observed, where sensitization towards these allergens at age 4- and 6 years exhibit the strongest association with rhinitis and eczema.

Discussion

Main Finding

This study revealed seven distinct patterns of allergic sensitization during preschool age by a NNS-PARAFAC cluster analysis. These seven novel patterns, grouped as *animal* pattern (cat/dog/horse), *other airborne* patterns (timothy grass/birch; molds; house dust mites), *mixed* pattern (peanut/wheat flour/mugwort), and *food* patterns (peanut/soy bean; egg/milk/wheat flour), were unique and reproducible in an independent cohort with high agreement. The patterns associated differentially to clinical disease development supporting that they represent biological meaningful sensitization phenotypes.

Strengths and Limitations of the Study

In our investigation of sIgE patterns during the first 6 years of life, the NNS-PARAFAC model allows usage of development over time as a third dimension in the analysis. The model “remembers” whether a child was sensitized at a prior time point, and takes that into account. Since prior sensitization affects your risk of being sensitized later, this information is an important contribution to the pattern analysis.

In contrast to the previous published methods for time course pattern recognition in the development of sensitization which are based on probabilistic hidden Markow models^{2,3}, the PARAFAC model constitutes a different framework, where, instead of estimating a probabilistic model, the PARAFAC model estimates a low dimensional representation of the observed data, with minimal distributional assumptions, in a similar fashion as principal component analysis for data organized in a matrix. This means that the quality of the resulting model is simply how good

a “picture” it is of the observed data, and how generalizable it is. The proposed clustering framework entails *soft clustering*, meaning that a single child can follow more than a single pattern, as well as an allergen can be described by several patterns. For example, elevated sIgE to peanut is described in both pattern 5 and pattern 6, and thereby contributes to two distinct groups of children; one with co-sensitization towards primarily mugwort and wheat flour (pattern 5) with primary data support at 6 years, and one with co-sensitization towards soybean, egg and wheat flour (pattern 6) with support from all time points. Further, the power of the PARAFAC model is in the ability to model higher ordered data structures without the need to unfold data into a matrix, and thereby abide to the nature of the data structure. In the specific application this has the upside of being able to estimate the time course directly, allowing for age dependent sensitization structures.

Association studies are typically conducted via individual comparison of all variables at all age-points with all outcomes^{4,18–23}. In the COPSAC study that amounts to 13 (allergens) x 4 (age-points) x 3 (outcomes) tests. This traditional approach has limited capability of bringing any novel biological insight due to: (1) the uncertainty in the original data is preserved, (2) the multiple testing issue, with a high probability of false discovery, and (3) the question answered by the statistical test, e.g. “is milk allergy at 1½yrs of age related asthma at 7 years?” is of limited biological relevance. The NNS-PARAFAC analysis circumvents these pitfalls by extracting only the systematic variation across both time and participants. Furthermore, as different types of allergens have similar mechanisms of sensitization, data driven grouping of these in the NNS-PARAFAC model reduces the number of tests without removing relevant

variation. Last, the biological question of interest is whether different sensitization patterns over time relates to clinical outcomes as opposed to single cross-sectional comparisons.

Both the COPSAC₂₀₀₀ and the BAMSE studies are strengthened by thorough data management and quality control procedures including external monitoring, online data collection into a dedicated database, and locking of data after data verification. A further advantage is that both studies longitudinally measured sIgE levels against the same diverse panel of common inhalant and food allergens. The almost exact replication of the seven latent sensitization patterns discovered in COPSAC₂₀₀₀ in the independent BAMSE cohort improves reliability of the findings.

A limitation of the study is the at-risk nature of the COPSAC₂₀₀₀ participants, i.e. children born to mothers with asthma, as this population may differ from unselected populations with regard to allergen exposures (house cleaning, pets in the house, etc.), which may impact sensitization in childhood^{24,20}. However, the replication of the patterns in the unselected BAMSE cohort argues against such inference. Further, replication is confined by the differences in study design between BAMSE and COPSAC₂₀₀₀, including age of assessment of sensitization (4 and 8 years in BAMSE and ½, 1 ½, 4, and 6 years in COPSAC₂₀₀₀), and age of assessment and diagnosis of clinical outcomes.

Interpretation

We discovered and verified seven distinct patterns of allergic sensitization during childhood, where groups of allergens follow a similar pattern over both time and participants. The grouping

was purely data driven and unbiased from prior knowledge on how allergens normally cluster (e.g. cross reactions) or which clinical outcomes normally associate with specific allergen sensitization.

The verification of the patterns in an independent cohort and the distinct association between the patterns and clinical outcomes are suggestive of biological meaningful sensitization trajectories.

Recent studies have also suggested the presence of several allergic sensitization patterns in childhood²⁻⁴. sIgE measurements from almost 13,000 children aged 3-17 years²⁵ found seven patterns very similar to ours (dog/cat/horse, house dust mites, molds, birch/apple, food/mugwort, timothy grass/rye and egg/milk). Although these data were cross-sectional and lacked the third dimension of temporal development in the statistical model, the similarity of the seven patterns supports our findings.

The association between eczema and all patterns with the exception of *other airborne* pattern 4 (house dust mite) aligns with previous studies showing association between eczema and allergic sensitization in general^{4,22,31}. Our finding that asthma was associated with *animal* pattern 1 (dog/cat/horse) in COPSAC and in children with asthmatic mothers in the BAMSE study is in line with other studies examining cat and dog sensitization independently^{4,18}, however, the study of Stoltz *et al* (2013)⁴ found the strength of association between dog and asthma to increase from age 1 year to 3 years, whereas our study suggest strongest association at age 1½ year. Many previous studies have reported that sensitization to house dust mite is related to asthma and wheezing^{2,4,18,21}. We only find a non-significant trend of association between pattern 4 (*D. pteronyssinus*) and asthma, and in the multivariate model, pattern 4 was not driving the association between sensitization and asthma either.

A British study of 1,053 unselected 8-year-old children (Manchester Asthma and Allergy Study – MAAS)² found a highly significant association between sensitization to house dust mites and asthma with an almost double effect size compared to our findings. We speculate that this discrepancy originates from COPSAC being a high-risk cohort with children from atopic mothers, as the association between pattern 4 and asthma based on all data in BAMSE yields results on par with the ones reported in MAAS.

When examining the patterns revealed in our study, there is a striking biological resemblance between the allergens with respect to components that constitute the allergens and the common protein-groups represented in each pattern. The allergens in pattern 1 (cat/dog/horse) comprise components that are both lipocalins (Can f 1, 2, 4, 6; Equ c 1, 2 and Fel d 4) and albumins (Can f 3, Equ c 3 and Fel d 2). Similarly, the allergens in pattern 2 (timothy grass/birch) comprise both profilins (Bet v 2 and Phl p 12) and polcalcins (Bet v 4 and Phl p 7). Allergens of pattern 5 (peanut/wheat flour/mugwort) comprise components that are nsLTPs (Art v 3; Ara h 9 and Tri a 14), pattern 6 (peanut/soybean) comprises both PR-10s (Ara h 8 and Gly m 4), legumins (Ara h 3 and Gly m 6), and vicilins (Ara h 1 and Gly m 5) and the allergens in pattern 7 (egg/milk/wheat flour) comprise components that are albumins (Bos d 6 and Gal d 5) except for wheat flour.

Although the components have biological resemblance in each pattern, this cannot alone explain the clustering of the allergens, since they need to cross-react to explain the patterns. The lipocalins that could be part of pattern 1 (cat/dog/horse) are known to cross-react²⁴, which is also apparent for polcalcins and profilins that could constitute pattern 2 (timothy grass/birch)^{26,27}. PR-10s are highly cross-reactive²⁸ which could explain the clustering in pattern 6 (peanut/soy bean). Some nsLTPs also cross-react²⁹, but the albumins of pattern 7 (egg/milk/wheat flour) are not

known to cross-react. This biological interpretation of the patterns is only speculative, and would need verification by component resolved diagnostics.

The association between rhinitis and all patterns with inhaled allergens was expected, whereas the association between rhinitis and *food* pattern 6 (peanut/soybean) might also be explained by cross-reactions as the Bet v 1 (from birch) related food allergen components in soy (Gly m 4)³⁰ and peanut (Ara h 8)²⁸ are all PR-10 proteins.

We propose the existence of a common epitope across patterns of different allergens, which most likely can be enlightened by examining allergen components. However, not all the patterns we found can be explained by cross-reactions and allergen components suggesting that other host, genetic, or environmental factors could also influence clustering of allergen specific sensitization during early childhood.

Conclusion

This study utilized a clustering method for three-dimensional data structures to extract distinct temporal patterns of sensitization while clustering allergens and children. These results suggest the presence of several distinct sensitization phenotypes during childhood defined by type and number of involved allergens, age at debut, and temporal development. The latent patterns showed differential associations with disease suggesting they represent biological meaningful and clinically relevant sensitization phenotypes.

Acknowledgements

We gratefully express our gratitude to the children and families of the COPSAC and BAMSE cohort studies for all their support and commitment. We acknowledge and appreciate the unique efforts of the COPSAC and BAMSE research teams, and Thermo Fisher Scientific Inc. for their dedicated work with the IgE analyses. We would also like to acknowledge Niklas Andersson, Institute of Environmental Medicine, Karolinska Institutet, for valuable statistical support.

Conflict of interest statement

The authors have no conflicts of interest relevant to this topic.
The corresponding author had full access to the data and had final responsibility for the decision to submit for publication.

References

1. Malby Schoos, A.-M. *et al.* Atopic endotype in childhood. *J. Allergy Clin. Immunol.* (2015). doi:10.1016/j.jaci.2015.10.004
2. Simpson, A. *et al.* Beyond Atopy. *Am J Respir Crit Care Med* **181**, 1200–1206 (2010).
3. Lazic, N. *et al.* Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* **68**, 764–770 (2013).
4. Stoltz, D. J. *et al.* Specific patterns of allergic sensitization in early childhood and asthma & rhinitis risk. *Clinical & Experimental Allergy* **43**, 233–241 (2013).
5. Kull, I. *et al.* Breast-feeding in relation to asthma, lung function, and sensitization in young schoolchildren. *Journal of Allergy and Clinical Immunology* **125**, 1013–1019 (2010).
6. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann. Allergy Asthma Immunol.* **93**, 381–389 (2004).
7. Bisgaard, H. *et al.* Childhood asthma after bacterial colonization of the airway in neonates. *N. Engl. J. Med* **357**, 1487–1495 (2007).
8. Bisgaard, H., Hermansen, M. N., Loland, L., Halkjaer, L. B. & Buchvald, F. Intermittent inhaled corticosteroids in infants with episodic wheezing. *N. Engl. J. Med* **354**, 1998–2005 (2006).
9. Wickman, M. *et al.* Childhood-to-adolescence evolution of IgE antibodies to pollens and plant foods in the BAMSE cohort. *Journal of Allergy and Clinical Immunology* **133**, 580–582 (2014).
10. Ballardini, N., Nilsson, C., Nilsson, M. & Lilja, G. ImmunoCAP Phadiatop Infant--a new blood test for detecting IgE sensitisation in children at 2 years of age. *Allergy* **61**, 337–343 (2006).
11. Braun-Fahrlander, C. *et al.* Validation of a rhinitis symptom questionnaire (ISAAC core questions) in a population of Swiss school children visiting the school health services. SCARPOL-team. Swiss Study on Childhood Allergy and Respiratory Symptom with respect to Air Pollution and Climate. International Study of Asthma and Allergies in Childhood. *Pediatr Allergy Immunol* **8**, 75–82 (1997).
12. Chawes, B. L. K. Upper and lower airway pathology in young children with allergic- and non-allergic rhinitis. *Dan Med Bull* **58**, B4278 (2011).
13. Chawes, B. L. K., Bønnelykke, K., Kreiner-Møller, E. & Bisgaard, H. Children with allergic and nonallergic rhinitis have a similar risk of asthma. *J. Allergy Clin. Immunol* **126**, 567–573–8 (2010).
14. Chawes, B. L. K., Kreiner-Møller, E. & Bisgaard, H. Objective assessments of allergic and nonallergic rhinitis in young children. *Allergy* **64**, 1547–1553 (2009).
15. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171 (1997).
16. Papalexakis, E. E., Sidiropoulos, N. & Bro, R. From k-means to higher-way co-

- clustering: multilinear decomposition with sparse latent factors. *Signal Processing, IEEE Transactions on* **61**, 493–506 (2013).
17. Allen, G. Sparse higher-order principal components analysis. in *International Conference on Artificial Intelligence and Statistics* 27–36 (2012).
 18. Brockow, I. *et al.* 3 Early Allergic Sensitizations and Their Relevance to Atopic Diseases in Children Aged 6 Years: Results of the GINI Study. *Journal of investigational allergology & clinical immunology* **19**, 180 (2009).
 19. Martinez, F. D. *et al.* Asthma and wheezing in the first six years of life. The Group Health Medical Associates. *N. Engl. J. Med* **332**, 133–138 (1995).
 20. Henderson, J. *et al.* Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. *Thorax* **63**, 974–980 (2008).
 21. Sears, M. R. *et al.* The relative risks of sensitivity to grass pollen, house dust mite and cat dander in the development of childhood asthma. *Clinical & Experimental Allergy* **19**, 419–424 (1989).
 22. Flohr, C. *et al.* Atopic Dermatitis and Disease Severity Are the Main Risk Factors for Food Sensitization in Exclusively Breastfed Infants. *J Invest Dermatol* (2013). doi:10.1038/jid.2013.298
 23. Lowe, A. J. *et al.* The temporal sequence of allergic sensitization and onset of infantile eczema. *Clinical & Experimental Allergy* **37**, 536–542 (2007).
 24. Eggleston, P. A. *et al.* Relationship of indoor allergen exposure to skin test sensitivity in inner-city children with asthma. *Journal of Allergy and Clinical Immunology* **102**, 563–570 (1998).
 25. Schmitz, R., Ellert, U., Kalckloesch, M., Dahm, S. & Thamm, M. Patterns of sensitization to inhalant and food allergens-findings from the German Health Interview and Examination Survey for Children and Adolescents. *International archives of allergy and immunology* **162**, 263–270 (2013).
 26. Verdino, P. *et al.* Three-dimensional structure of the cross-reactive pollen allergen Che a 3: visualizing cross-reactivity on the molecular surfaces of weed, grass, and tree pollen allergens. *The Journal of Immunology* **180**, 2313–2321 (2008).
 27. Casquete-Román, E. *et al.* Profilin cross-reactive panallergen causes latex sensitization in the pediatric population allergic to pollen. *Annals of Allergy, Asthma & Immunology* **109**, 215–219 (2012).
 28. Mittag, D. *et al.* Ara h 8, a Bet v 1-homologous allergen from peanut, is a major allergen in patients with combined birch pollen and peanut allergy. *Journal of Allergy and Clinical Immunology* **114**, 1410–1417 (2004).
 29. Jimenez-Lopez, J. C., Gachomo, E. W., Ariyo, O. A., Baba-Moussa, L. & Kotchoni, S. O. Specific conformational epitope features of pathogenesis-related proteins mediating cross-reactivity between pollen and food allergens. *Molecular biology reports* **39**, 123–130 (2012).
 30. Ballmer-Weber, B. K. & Vieths, S. Soy allergy in perspective. *Current opinion in allergy and clinical immunology* **8**, 270–275 (2008).

31. Carlsten, C. *et al.* Atopic dermatitis in a high-risk cohort: natural history, associated allergic outcomes, and risk factors. *Annals of Allergy, Asthma & Immunology* **110**, 24–28 (2013).

Figures

Figure 1: Positive registrations in relative frequency of the 13 allergens at the four age-points

Figure 2: Pairwise comparison of latent patterns between the COPSAC₂₀₀₀ data (grey) and the BAMSE data (black), showing groups of specific allergen patterns that follow a similar profile in both time and participants. **A** (upper panel): unit norm sensitization pattern loadings (x-axis describes loading magnitude). **B** (lower panel): time-development pattern loadings (y-axis describes loading magnitude).

Figure 3: Association between sensitization patterns and clinical outcomes for COPSAC₂₀₀₀ data (n = 398) (black) and for children of asthmatic mothers for BAMSE data (n = 282) at 4 years (light grey) and at 8 years (dark grey). For asthma and rhinitis the comparison is done with logistic regression and for eczema in COPSAC₂₀₀₀ by Cox regression and in BAMSE by logistic regression.

Figure 4: Age dependent associations between patterns and clinical outcomes shown as OR / HR (black line) with CI_{0.95} (dotted line) in COPSAC₂₀₀₀. Some estimates for pattern 1-5 at age ½y and 1½y are omitted due too low number of positive sensitization registrations for the particular pattern. Included are the association results based on the clustering from all time points (black star with error bars).

Tables

Table 1: Baseline characteristics in the COPSAC₂₀₀₀ and BAMSE cohorts.

	COPSAC	BAMSE ast ⁺	BAMSE all
	N=398	N=282	N=3045
Gender, male (%)	199 (50%)	156 (55%)	1554 (51%)
Smoking (last trimester), yes (%) [§]	59 (15%)	31 (11%)	280 (9.2%)
SOCIOECONOMIC FACTORS			
Household income*, High/Average/Low (% Low)	90/183/115 (30%)	NA	NA
Mother's education in COPSAC, University/Medium/At most college (% At most college)	52/100/246 (62%)	NA	NA
Mother's education in BAMSE, Higher and Medium white/Lower white/Blue collar worker (% At most Blue collar worker)	NA	126/57/65 (26%)	1408/683/676 (24%)
ATOPIC PREDISPOSITION			
Mother			
Asthma, yes (%)	398 (100%)	282 (100%)	282 (9.3%)
Allergic rhinitis, yes (%)	310 (78%)	181 (64%)	550 (18%)
Atopic dermatitis, yes (%)	192 (48%)	92 (33%)	521 (17%)
ATOPIC PREDISPOSITION			
Father			
Asthma, yes (%)	74 (19%)	18 (6.5%)	239 (7.9%)
Allergic rhinitis, yes (%)	142 (37%)	44 (16%)	518 (17%)
Atopic dermatitis, yes (%)	54 (14%)	25 (9.1%)	362 (12%)
ALLERGEN EXPOSURES			
Dog in the home, yes (%) [€]	52 (13%)	17 (6.0%)	147 (4.8%)
Cat in the home, yes (%) [€]	60 (15%)	20 (7.1%)	296 (9.7%)

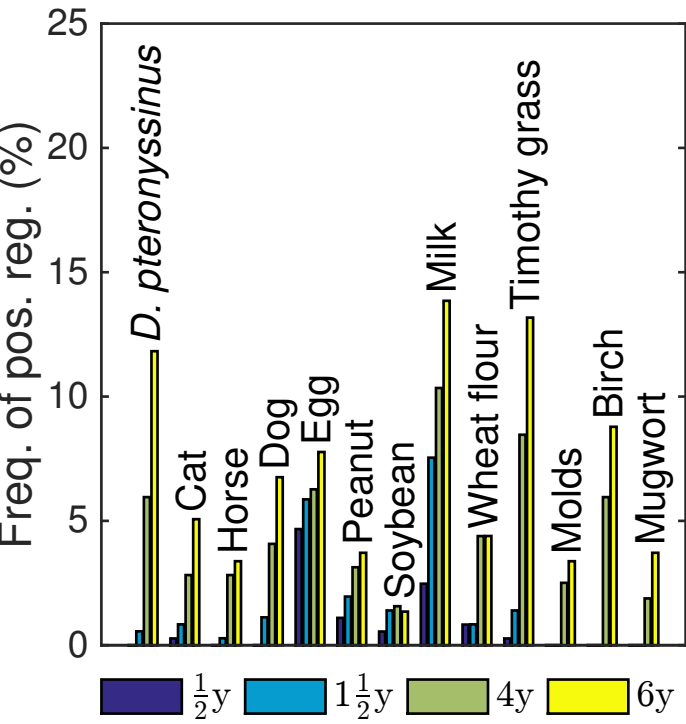
⁺ Children born to mothers with asthma in the BAMSE cohort

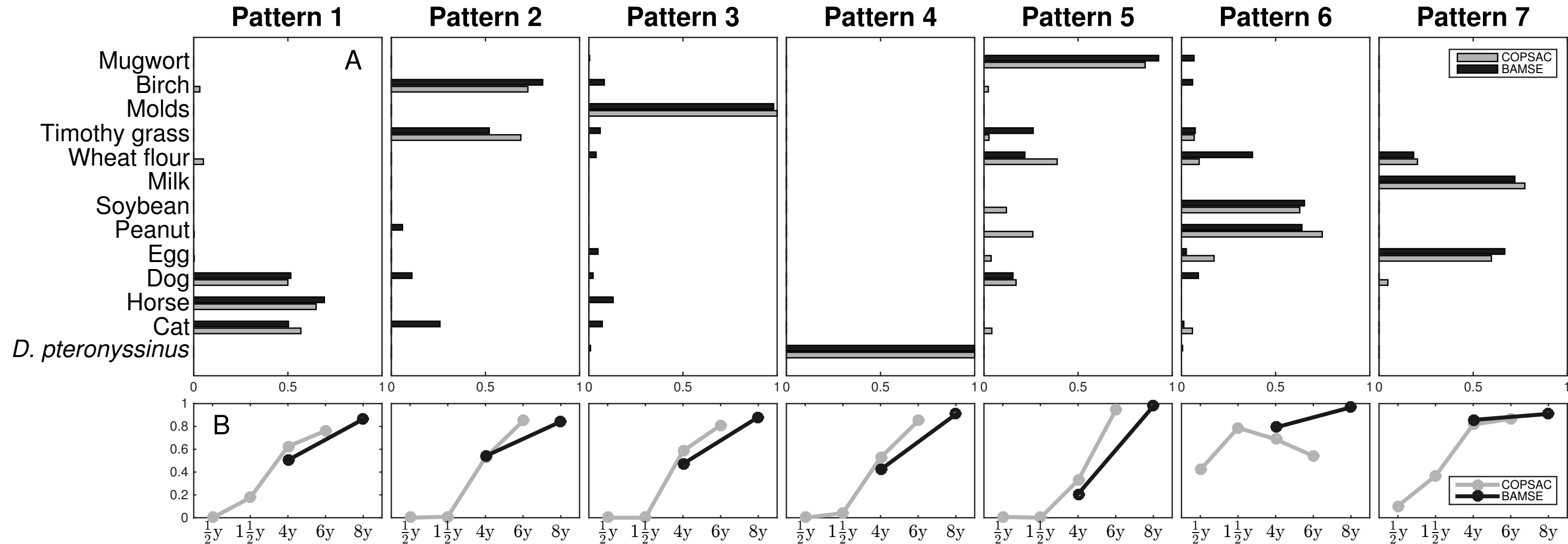
* High is defined as a total income > €80,000, average from €50,000 to 80,000, low is < €50,000

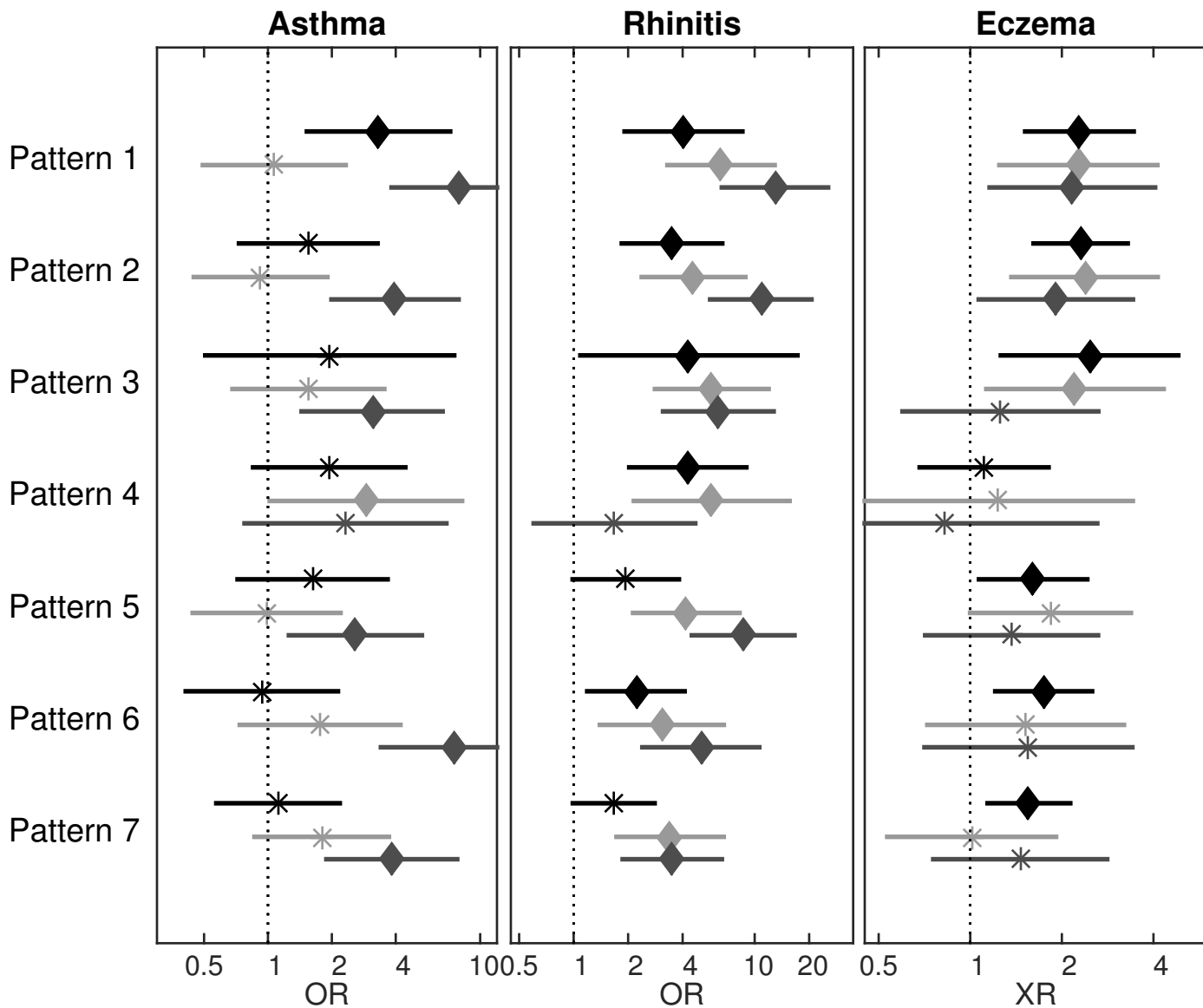
^ Lower white=2 years or less at a higher education, Medium and High white=At least 3 years at a higher education

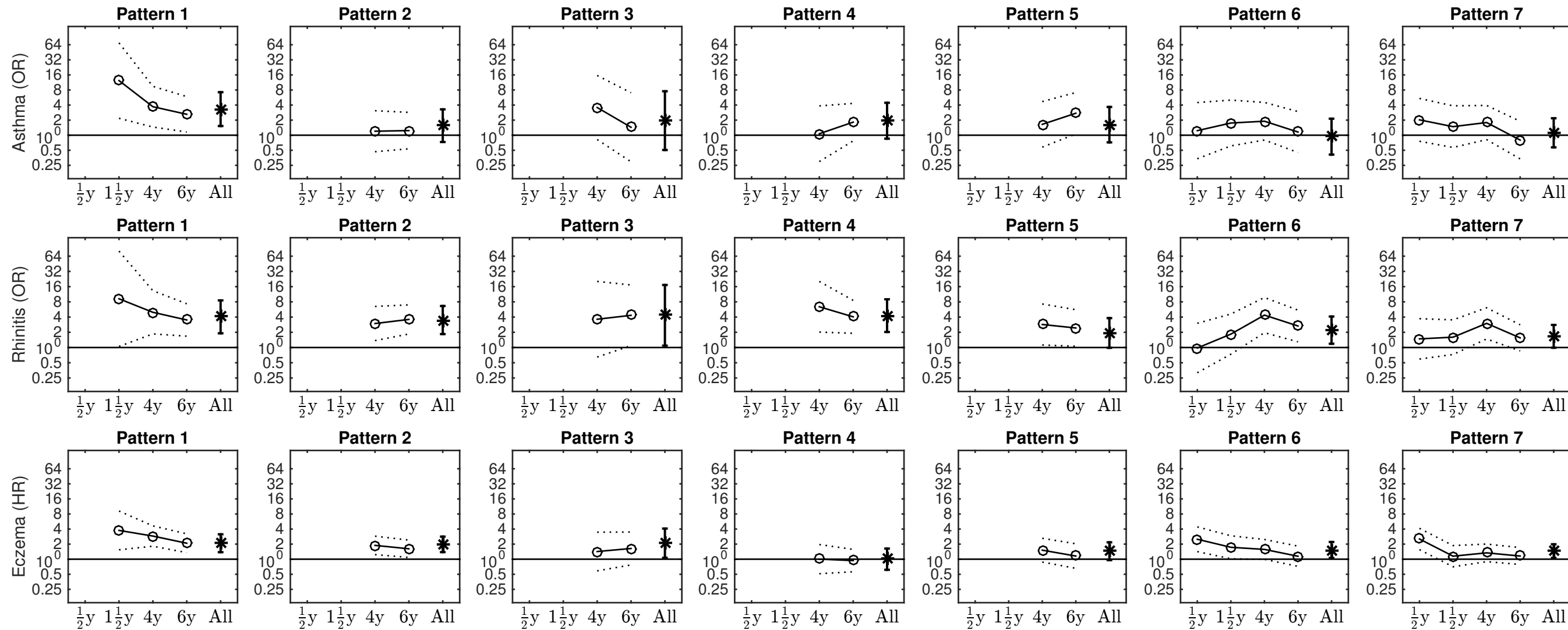
[§] ≥ 1 cigarette/week

[€] First year of life









1 **Novel Sensitization Trajectories in Childhood revealed by a Cluster**

2 **Analysis**

3 Ann-Marie M Schoos, MD, PhD; Bo L Chawes, MD, PhD, DMSc; Erik Melén, MD,
4 PhD; Anna Bergström, PhD; Inger Kull, PhD; Magnus Wickman, MD, PhD; Klaus
5 Bønnelykke, MD, PhD; Hans Bisgaard, MD, DMSc; Morten A Rasmussen, PhD

6 Online Repository

7

8

9 **Methods**

10 *Clinical Outcomes in BAMSE*

11 Asthma in BAMSE was diagnosed by questionnaire as more than 3 episodes of
12 wheeze in the last 12 months and/or; at least 1 episode of wheeze in the last 12
13 months, combined with prescription of inhaled steroids for symptoms of asthma for
14 both the 4- and the 8yrs data.

15
16 Rhinitis in BAMSE was diagnosed by questionnaire as symptoms from eyes/nose
17 (suspected or evident) after exposure to furred pets or pollen (after age 2yrs) and/or
18 doctor's diagnosis of allergic rhinitis after 2 years and up to age 4 years for 4yrs data,
19 and symptoms of sneezing, a runny or blocked nose, or itchy, red and watery eyes
20 after exposure to furred pets or pollen (from age 4 years) and/or doctor's diagnosis of
21 allergic rhinitis from 4 years and up to age 8 years for 8yrs data.

22 Eczema in BAMSE was diagnosed by questionnaire as dry skin in combination with
23 itchy rash for 2 weeks or more at typical localization (face or arms/legs extension
24 surfaces or arms/legs flexures or wrists/ankles flexures) in the last 12 months and/or
25 doctor's diagnosis of eczema after age 2- or 4 years up to age 4- or 8 years for
26 diagnosis at 4yrs and 8yrs respectively.

27

28 Data Analysis

29 *Nonnegative Sparse PARAllel FACtor analysis*

30 Let $\mathbf{X}$ be a tensor of size (n, m, p) , where n is the number of samples (children), m is
 31 the number of time points and p is the number of allergens. The Nonnegative Sparse
 32 PARAllel FACtor analysis (NNS-PARAFAC) method aims at approximating the
 33 observed data ($\mathbf{X}$) with a sum of rank-one tensors in a similar fashion as principal
 34 component analysis for data organized in a matrix. That leads to the main objective in
 35 equation E1

$$36 \min_{\mathbf{A}, \mathbf{B}, \mathbf{C}} \left(\|\tilde{\mathbf{X}} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F^2 \right) \quad (\text{E1})$$

37 Where $\tilde{\mathbf{X}} (n, mp)$ is an unfolded version of $\mathbf{X}$, $\mathbf{A} (n, k)$ is a factor matrix related to the
 38 samples, $\mathbf{B} (m, k)$ is a factor matrix related to the time points and $\mathbf{C} (p, k)$ is a factor
 39 matrix related to the allergens, $\odot$ is the so called Khatri-Rao product of two matrices
 40 and $\|\cdot\|_F^2$ is the squared frobenius norm (sums of squares across all the matrix
 41 elements). The main objective (E1) can be modified to fulfill certain constraints on
 42 the parameters ($\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$). In our case, that amounts to non-negativity for all factor
 43 matrices, as well as a L_1 sparsity constraint on the columns of $\mathbf{C}$. This leads to the
 44 constrained problem in equation E2.

$$45 \min_{\mathbf{A}, \mathbf{B}, \mathbf{C}} \left(\|\tilde{\mathbf{X}} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F^2 \right)$$

subject to (E2)

$$A_{ij} \geq 0 \forall i, j; B_{ij} \geq 0 \forall i, j; C_{ij} \geq 0 \forall i, j$$

$$|\mathbf{C}_i|_1 \leq \kappa_i \forall i = 1, \dots, k$$

46 Where $|C_{i.}|_1 = \sum_{j=1}^p |C_{ij}|$ is the so-called L_1 norm or Manhattan norm and κ_i is a
 47 scalar tuning parameter.

48

49 *Factorization and bi-/co-clustering*

50 The incorporation of sparsity via non-negativity and the L_1 norm constraint yields a
 51 sparse solution. The samples corresponding to the component wise non-zero elements
 52 of **A** forms a cluster, as these samples actively support estimation of the data (**X**),
 53 while the samples with elements equal to zero does not, and are therefore acting as a
 54 reference cluster. Further, as each component is made up from ingredients reflecting
 55 both samples (**A**), time-points (**B**) and allergens (**C**), where the latter is also sparse,
 56 clustering is obtained on both samples and allergens simultaneously leading to the
 57 notion of co-clustering, where a subset of the samples (children) are active (non-
 58 negative score values in **A**) with respect to a subset of the allergens (non-negative
 59 loading values in **C**).

60

61 *Algorithm*

62 Estimation of the factor matrices is done by alternating least squares, where each
 63 factor matrix (say **A**) is updated conditioning on the current estimate of the other
 64 factor matrices (say **B** and **C**). This procedure is repeated until the parameter
 65 estimates remain stable. In contrast to other alternating procedures, the algorithm
 66 optimizes the entire set of k components, as opposed to one component at a time
 67 followed by deflation of the data. The main objective (E1) is not convex, and as such
 68 a solution might not necessarily be the global optimum. Therefore a range of models

69 with different initial parameter guesses are produced and compared. The one with the
70 lowest error that is found several times from 100 random starts is considered as the
71 solution. The algorithm is available through www.models.life.ku.dk/algorithms

72 *Prediction of a new sample*

73 Let $\mathbf{X}^*$ be a new sample (a matrix of size (m, p) , where m is the number of time points
74 p is the number of allergens), and let $\mathbf{B}$ (m, k) and $\mathbf{C}$ (p, k) be estimated factor
75 matrices. Then a corresponding new set of parameters $\mathbf{a}^*$ ($1, k$) is estimated by
76 solving the following objective.

$$\min_{\mathbf{a}^*} \left(\|\widetilde{\mathbf{X}}^* - \mathbf{a}^*(\mathbf{C} \odot \mathbf{B})^T\|_F^2 \right)$$

77 subject to (E3)

$$\mathbf{a}^*_i \geq 0 \quad \forall i = 1, \dots, k$$

78 Where $\widetilde{\mathbf{X}}^*$ ($1, mp$) is an unfolded version of $\mathbf{X}^*$. Similarly, if $\mathbf{B}$ (the time profile factor
79 matrix) is not fixed, and hence should be estimated based on the new data, this matrix
80 is left as variable within the objective. The algorithm for this procedure follows from
81 the above, excluding update of fixed factor matrices.

82 *Data Preprocessing*

83 The data are initially scaled across allergens by the inverse of the square root of the
84 sum of the number of positive registrations. This scaling up weights allergens with
85 few positive registrations and down weights those with many positive registrations.

86 *Model tuning*

87 This model has tuning parameters in relation to the sparsity constraints (κ_i) and the
88 number of components. The sparsity constraints are estimated by Bayes Information

Criterion for each component separately, balancing fit (main loss as in equation E1) and degrees of freedom, which is set to the number of non-zero parameters. The number of components is selected based on comparing models with varying numbers of components, and judging i) additional variation explained from inclusion more components as well as ii) component interpretability.

Post hoc clustering at each time point

The NNS-PARAFAC model reveals clusters of children covering the entire age span. However, as we are interested in the age where the sensitization penetrates with respect to the clinical outcome, the model is used to cluster the children at each time point. This is done by estimating the objective in equation E3.

$$\min_{\mathbf{A}} (\|\mathbf{X}_t - \mathbf{A}_t \mathbf{C}^T\|_F^2)$$

subject to (E3)

$$A_{ij} \geq 0 \forall i, j$$

Where $\mathbf{X}_t$ ($t = \frac{1}{2}y, 1\frac{1}{2}y, 4y, 6y$) corresponds to data from the individual time points, $\mathbf{A}_t$ is the corresponding factor matrix and $\mathbf{C}$ is the estimated factor matrix from the NNS-PARAFAC model as described in equation E2. I.e. this model estimates the clustering of the children in accordance with the predefined allergen profiles.

Script

Below is shown the code for preprocessing the data, and running the NNS-PARAFAC model with multiple random starts. Be aware that this script works generically for both datasets.

```

109 % INPUT: X (n,p,q) - threeway data structure with dicotomized (0,1) sIgE
110 registrations
111 % for q = 13 allergenes, across p = 4 timepoints with n = (398 / 3051)
112 % children
113
114 % 1) scale data
115 % Scale to equal sum across all variables!
116 sumX = nansum(reshape(permute(X.data,[3 1
117 2]),size(X,3),size(X,2)*size(X,1))');
118 Xsc = X;
119 for i=1:size(X,3);
120     scl(i) = (1/sqrt(sumX(i)));
121     Xsc.data(:, :, i) = X.data(:, :, i)*scl(i);
122 end
123
124 % 2) model settings
125 [I,J,K] = size(X);
126 sumabs = [sqrt(I) sqrt(J) -1];
127 nonneg = ones(3,1);
128 orth = zeros(3,1);
129
130 % 3) multiple starts
131 for i=1:100;
132     m = nnsparafac(Xsc.data, 7, sumabs, nonneg, orth);
133     M{i} = m;
134     SSE(i) = m.sse;
135 end
136
137 % 4) Check model convergens
138 hist(SSE)
139
140 % 5) select best model
141 [~, id] = min(SSE);
142 model = M{id};
143
144
145

```

146 *Verification*

147 The NNS-PARAFAC model is an unsupervised single block model, meaning that the
 148 model is build without focus on predicting an outcome, and is consequently not prone
 149 to over fitting. In this application, we utilize the fact that the NNS-PAFARAC model
 150 is mathematically unique (up to component permutation) which implies that two
 151 models obtaining the exact same loss in equation E1, consequently has the same
 152 parameter estimates. We use this to validate the allergen patterns by estimation of a
 153 model on independent data (BAMSE) followed by a direct comparison of the
 154 parameter estimates. Further, the comparison with clinical endpoints is done using
 155 similar methodology between the two cohorts. Additionally, we also used the
 156 estimated model based on the COPSAC₂₀₀₀ data to individually predict the clustering
 157 of each child from the BAMSE data by the method described above in *Prediction of a*

new sample and reference this clustering towards the clustering based on the BAMSE data alone using Cohen's kappa as measure of agreement.

Robustness testing

In order to access the robustness of the pattern results towards random and specific changes in study population as well as compliance for the sensitization data three strategies were pursued. I) Based on maternal predisposition towards asthma, rhinitis and eczema the cohorts were split into all possible combinations revealing four groups in COPSAC and eight groups in BAMSE. Each of these stratum was used for building a seven-component pattern model with the same settings as for the model based on all data. These models were compared with the model based on all data. II) The results were tested for robustness towards random perturbations of the study population by applying a bootstrap scheme for random selection of children followed by model estimation and comparison with model build on the entire dataset. III) In the COPSAC cohort $n = 252$ children has complete follow up with respect to sensitization data. Based on these samples a model was estimated and compared with a model based on the entire dataset. The comparisons were conducted using pairwise Pearson's correlations between allergen loadings.

Multivariate regression

In addition to associating the extracted sensitization patterns one-by-one towards the clinical outcomes, a multiple regression model is investigated. Initially, all the seven dichotomized pattern scores are used in one model. This model is modified by backward elimination to only include the significant ($p < 0.05$) patterns.

Results

Compliance

Of the 398 children included in this study, 25 (6.3%) only contributed with data from a single time point, while 51 (12.8%) and 70 (17.6%) contributed with data from two- or three time points respectively. 252 (63.3%) children have complete follow up from 6 month to 6 years with respect to specific IgE data. While the longitudinal trends demands longitudinal data, clustering of the allergens obtain support from all measured information including cross sectional data points. Therefor all data (n = 398) were included in the further analysis.

Specific IgE data

The registered levels of sIgE in the COPSAC₂₀₀₀ cohort are shown in **Figure E1**. In total 4653 samples (out of 17,420 registrations) (27%) obtained a value larger than zero.

Abundance of allergens in COPSAC₂₀₀₀ and BAMSE

In an allergen wise comparison of abundances between the COPSAC₂₀₀₀ and BAMSE cohorts at age 4yrs, the only significant differences were sensitization against *D. pteronyssinus* (higher relative abundance in COPSAC₂₀₀₀, $p < 0.0001$), cat (lower relative abundance in COPSAC₂₀₀₀, $p = 0.02$) and molds (higher relative abundance in COPSAC₂₀₀₀, $p = 0.01$) (**Figure E2**).

Abundance of Sensitization Patterns in COPSAC₂₀₀₀ and BAMSE

The abundance of the seven sensitization patterns extracted from COPSAC₂₀₀₀ and BAMSE are shown in **Figure E3**. In general, a higher abundance is registered in BAMSE, especially when stratifying on maternal asthma. Pattern 3, 5 and 7 in

COPSAC₂₀₀₀ and BAMSE respectively are, however, not significantly different ($p > 0.06$) in abundance when stratifying on maternal asthma.

Pattern verification in BAMSE

Table E1 shows the Cohen's kappa agreement for each of the seven patterns between clustering of the children based on a model build on COPSAC₂₀₀₀ data and a model build on BAMSE data. Pattern 1, 2, 4, 5 and 7 are in good agreement between the two clustering procedures ($\kappa > 0.74$), whereas especially pattern 3 (*Molds*) is in poor agreement ($\kappa = 0.36$). This lack of agreement derives from the BAMSE model having a total of nine allergens active for this component, as opposed to the COPSAC₂₀₀₀ model with only a single allergen (*Molds*) which results in a 4 fold higher number of cases.

Robustness of Sensitization Patterns

The effect of maternal predisposition (asthma, eczema and rhinitis) was investigated on the sensitization prevalence, as well as the stability of the patterns. **Figure E4** shows the sensitization frequency as function of allergen, and maternal predisposition. Obviously, predisposition boosts the prevalence of sensitization, where especially dermatitis promotes sensitization. **Figure E5** shows the pattern agreement between a model estimated on the entire data (within each cohort) and models stratified for maternal predisposition. In general, the agreement is good (median $r = 0.75$ – BAMSE, median $r = 0.65$ – COPSAC), however, for pattern 3 in BAMSE (dominated by molds) there is in general poor agreement. Molds is also one of the least frequent allergens the children were sensitized to (~1%), and reducing sample size therefore severely reduces data support for capturing this pattern. In conclusion, the sensitization patterns seem robust towards changes in population, such that

maternal predisposition induce a change in sensitization prevalence, but not in the patterns of co-occurrence. The only requirement for model estimation is a large enough sample size in order to have data support for the least common allergens.

Figure E6 shows the robustness of the sensitization patterns towards random changes in the study population, as pattern wise correlations between the bootstrapped models and original model estimated on both cohorts. The COPSAC model reveals on average the model is very robust ($r_{\text{median}} > 0.94$ for all patterns). However the estimation of pattern 2, 5 and 6 are prone to changes in population as up to 10% of the bootstrapped models obtain $r < 0.8$ for these patterns. The larger BAMSE cohort is very robust in terms of random population changes as $r > 0.90$ for all patterns, all repetitions.

Figure E7 shows the robustness of the sensitization patterns from an analysis only including children with complete follow up from 6 months to 6 years with respect to sIgE data ($n = 252$). These results indicate robustness towards changes in inclusion criterion. The association results between the seven patterns and the three clinical outcomes are likewise robust towards this change in inclusion criteria (results not shown).

Multivariate regression models

Based on the COPSAC₂₀₀₀ data, a multivariate model with asthma at seven years was reduced to only include pattern 1. A multivariate model predicting rhinitis at seven years consisted of two significant patterns, namely pattern 2 ($p = 0.03$) and pattern 4 ($p = 0.005$). For time to onset of eczema, the significant patterns were found to be pattern 1 ($p = 0.02$) and pattern 2 ($p = 0.006$) (see **Table E2-4**). Results based on BAMSE data confirms pattern 1 as important for predicting asthma and eczema, and

pattern 4, and to some extent pattern 2, relevant for predicting rhinitis. However, especially the BAMSE models based on all available data ($n = 2604$) suggest additional patterns to be important.

Adjustment for comorbidity

Of the three outcomes, only asthma and rhinitis are associated ($p = 0.01$) and hence possibly confounds one another in the association analysis towards the sensitization patterns. When adjusting the association towards rhinitis for asthma the results are unaffected, whereas the pattern-wise association with asthma attenuates when adjusting for rhinitis (see **Table E5** – based on COPSAC₂₀₀₀).

Figures

Figure E1: Registered levels of sIgE in the COPSAC₂₀₀₀ cohort (log transformed). The green line indicates the dichotomization threshold 0.35 kU_A/L.

Figure E2: Relative abundance of 13 allergens matched at 4 years of age from the COPSAC₂₀₀₀ cohort ($n=296$) and the BAMSE cohort ($n=2604$).

Figure E3: Abundance (relative number of children) in each pattern. BAMSE (all) refer to the total cohort ($n=3051$) and BAMSE (maternal asthma) refer to the strata with children of asthmatic mothers ($n=282$)

Figure E4: Prevalence of sensitization as a function of maternal predisposition. The frequencies are any positive registration across the entire time course divided by the total number of children within each subgroup.

Figure E5: Pattern agreement (Pearson's correlations) between a model build on all samples and models build on samples defined by maternal predisposition. Numbers in parenthesis refer to total n within strata and number of samples with at least a single positive (>0) sIgE registration.

Figure E6: Bootstrap results based on the COPSAC₂₀₀₀ data. Each of 100 bootstrapped models is compared to the model based on all data, by assessing the correlation between the individual patterns.

Figure E7: Comparison of COPSAC₂₀₀₀ NNS-PARAFAC models based on children with complete follow up (black - n = 252) and children with at least a single sensitization data point (grey - n = 398). **A:** Allergen profiles-, and **B:** age trajectories for the seven patterns.

Figure E8: Association between sensitization patterns and clinical outcomes for COPSAC₂₀₀₀ data (n = 398) (black) and BAMSE (n = 3051) data at 4 years (light grey) and at 8 years (dark grey). For eczema the effect size is registered as hazard ratios in COPSAC₂₀₀₀ and odds ratios in BAMSE.

297

298 **Figure E9:** Size of clusters at each time point, as well as all time points.

ACCEPTED MANUSCRIPT

299 Tables

300 **Table E1:** Agreement (Cohen's kappa $\pm$ SE) between clustering of children in
 301 BAMSE based on either COPSAC₂₀₀₀ model or BAMSE model.

	N (COPSAC ₂₀₀₀ model)	N (BAMSE model)	Cohens kappa ($\kappa \pm$ SE)
Pattern 1	505 (17%)	429 (14%)	0.90 $\pm$ 0.01
Pattern 2	577 (19%)	649 (21%)	0.92 $\pm$ 0.01
Pattern 3	69 (2%)	289 (9%)	0.36 $\pm$ 0.04
Pattern 4	121 (4%)	121 (4%)	1.00 $\pm$ 0.00
Pattern 5	442 (14%)	463 (15%)	0.74 $\pm$ 0.02
Pattern 6	504 (17%)	345 (11%)	0.58 $\pm$ 0.02
Pattern 7	489 (16%)	556 (18%)	0.87 $\pm$ 0.01

302

303 **Table E2:** Multivariate models (OR [$CI_{0.025}$; $CI_{0.975}$]) of **asthma** in COPSAC₂₀₀₀,
 304 children of asthmatic mothers in BAMSE, and all children in BAMSE estimated by
 305 backward elimination.

	COPSAC ₂₀₀₀	BAMSE			
	Asthmatic mothers (n = 398)	Asthmatic mothers (n = 282)		All (n = 2604)	
	Age 7 yrs	Age 4 yrs	Age 8 yrs	Age 4 yrs	Age 8 yrs
Pattern 1	3.32 [1.53;7.20]	ns	5.28 [2.42;11.50]	2.15 [1.49;3.10]	3.91 [2.71;5.62]
Pattern 2	ns	ns	ns	ns	ns
Pattern 3	ns	ns	3.76 [1.55;9.11]	1.69 [1.14;2.51]	2.65 [1.80;3.89]
Pattern 4	ns	ns	ns	ns	ns
Pattern 5	ns	ns	ns	ns	ns
Pattern 6	ns	ns	ns	ns	ns
Pattern 7	ns	2.89 [1.02;8.18]	ns	2.25 [1.38;3.66]	1.67 [1.00;2.79]

306 Pattern 1= cat/dog/horse, pattern 2= timothy grass/birch, pattern 3= molds, pattern 4=
 307 house dust mites, pattern 5= peanut/wheat flour/mugwort, pattern 6= peanut/soybean,
 308 pattern 7= egg/milk/wheat flour.

309 ns = marginal non-significant ($p > 0.05$)

310

Table E3: Multivariate models (OR [CI_{0.025}; CI_{0.975}]) of **rhinitis** in COPSAC₂₀₀₀, children of asthmatic mothers in BAMSE, and all children in BAMSE estimated by backward elimination.

	COPSAC ₂₀₀₀	BAMSE			
	Asthmatic mothers (n = 398)	Asthmatic mothers (n = 282)		All (n = 2604)	
	Age 7 yrs	Age 4 yrs	Age 8 yrs	Age 4 yrs	Age 8 yrs
Pattern 1	ns	4.18 [1.85;9.42]	7.19 [3.28;15.76]	2.98 [1.99;4.45]	3.81 [2.69;5.40]
Pattern 2	2.92 [1.51;5.64]	ns	3.05 [1.36;6.80]	ns	1.44 [1.03;2.02]
Pattern 3	ns	ns	ns	1.61 [1.16;2.24]	1.42 [1.01;1.99]
Pattern 4	3.49 [1.62;7.51]	ns	ns	1.68 [1.15;2.47]	7.06 [4.83;10.33]
Pattern 5	ns	ns	ns	ns	ns
Pattern 6	ns	2.47 [1.03;5.95]	ns	2.00 [1.43;2.81]	1.75 [1.22;2.52]
Pattern 7	ns	ns	ns	ns	0.49 [0.30;0.81]

Pattern 1= cat/dog/horse, pattern 2= timothy grass/birch, pattern 3= molds, pattern 4= house dust mites, pattern 5= peanut/wheat flour/mugwort, pattern 6= peanut/soybean, pattern 7= egg/milk/wheat flour.

ns = marginal non-significant ($p > 0.05$)

Table E4: Multivariate models (HR / OR [CI_{0.025}; CI_{0.975}]) of **eczema** in COPSAC₂₀₀₀, children of asthmatic mothers in BAMSE, and all children in BAMSE estimated by backward elimination in a cox model (COPSAC₂₀₀₀) and logistic regression model (BAMSE).

	COPSAC ₂₀₀₀	BAMSE			
	Asthmatic mothers (n = 398)	Asthmatic mothers (n = 282)		All (n = 2604)	
	Cox model	Age 4 yrs	Age 8 yrs	Age 4 yrs	Age 8 yrs
Pattern 1	1.60 [1.00;2.55]	ns	2.17 [1.16;4.04]	1.58 [1.14;2.19]	1.74 [1.22;2.50]
Pattern 2	1.66 [1.11;2.50]	ns	ns	ns	ns
Pattern 3	ns	ns	ns	ns	1.49 [1.09;2.04]
Pattern 4	ns	2.37 [1.37;4.12]	ns	1.89 [1.43;2.50]	1.49 [1.08;2.06]

Pattern 5	ns	ns	ns	ns	ns
Pattern 6	ns	ns	ns	1.56 [1.16;2.10]	ns
Pattern 7	ns	ns	ns	ns	ns

322 Pattern 1= cat/dog/horse, pattern 2= timothy grass/birch, pattern 3= molds, pattern 4=
323 house dust mites, pattern 5= peanut/wheat flour/mugwort, pattern 6= peanut/soybean,
324 pattern 7= egg/milk/wheat flour.

325 ns = marginal non-significant ($p > 0.05$)

326

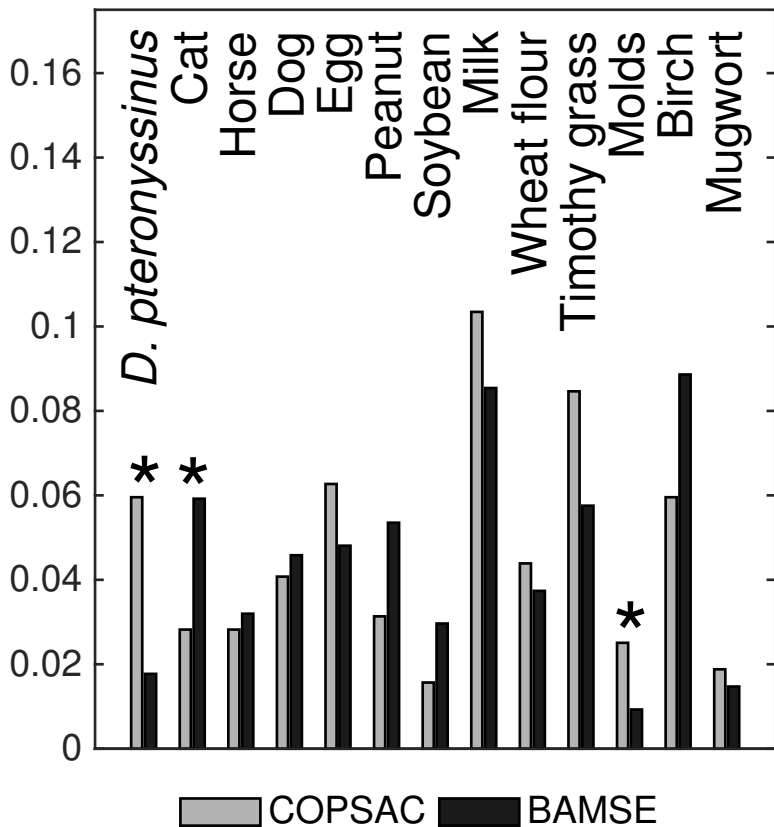
327 **Table E5:** Association between the different patterns of sensitization and clinical
328 outcomes in COPSAC₂₀₀₀. The outcome by pattern frequency is relative to the number
329 of children within the sensitization pattern with outcome registration.

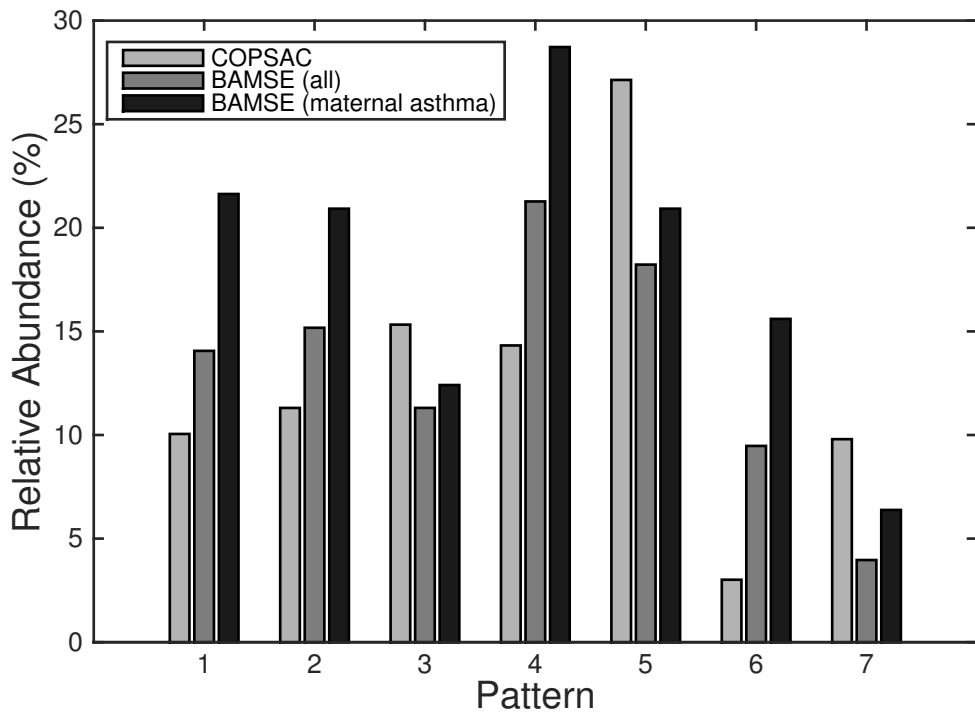
		PATTERNS [§]						
		Entire cohort	1 <i>Animal</i>	2 <i>Other airborne</i>	3 <i>Other airborne</i>	4 <i>Other airborne</i>	5 <i>Mixed</i>	6 <i>Food</i>
N (%)		398	40 (10%)	57 (14%)	12 (3%)	39 (10%)	45 (11%)	61 (15%)
Asthma 7yrs n (%)		47 (14%)	12 (32%)	11 (20%)	3 (25%)	9 (24%)	9 (21%)	8 (14%)
OR, CI			3.3 [1.5;7.2]	1.5 [0.7;3.3]	2.0 [0.5;7.5]	1.9 [0.9;4.4]	1.6 [0.7; 3.7]	0.9 [0.4; 2.1]
aOR*, CI			2.2 [0.9;5.2]	0.9 [0.4;2.1]	0.9 [0.2;4.7]	1.1 [0.4;2.8]	1.5 [0.7;3.7]	0.7 [0.3;1.7]
Rhinitis 7yrs n (%)		105 (36%)	23 (66%)	30 (61%)	7 (70%)	24 (67%)	20 (50%)	26 (52%)
OR, CI			4.0 [1.9;8.6]	3.5 [1.8;6.6]	4.3 [1.1;17]	4.3 [2.0;9.0]	1.9 [1.0;3.8]	2.2 [1.2;4.1]
aOR^, CI			3.7 [1.7;8.1]	3.7 [1.9;7.2]	4.3 [1.0;18]	4.1 [1.9;8.7]	1.9 [0.9;4.0]	2.3 [1.2;4.4]
Eczema 7 yrs n (%)		177 (52%)	27 (71%)	39 (70%)	9 (75%)	18 (50%)	27 (64%)	36 (62%)
								59 (60%)

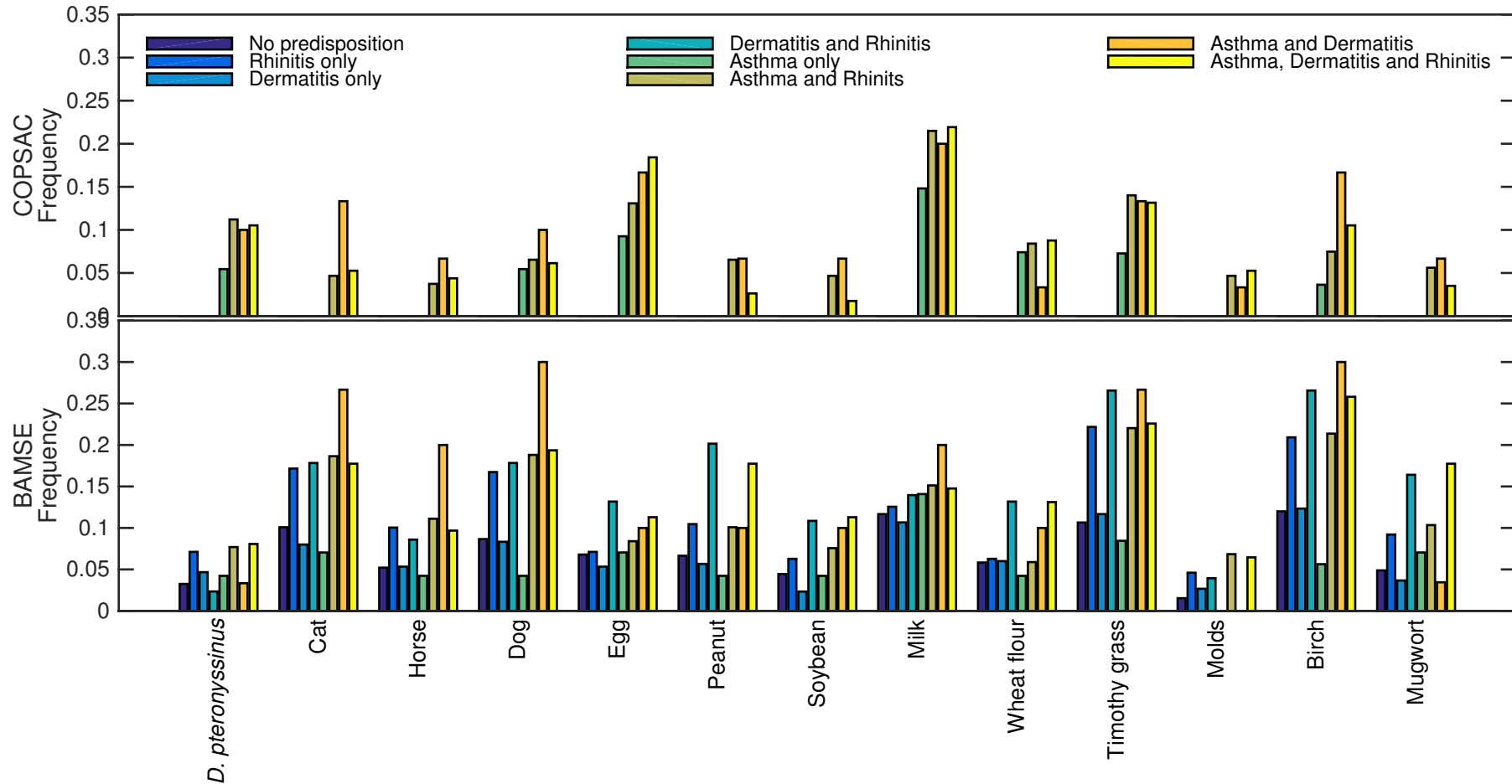
HR, CI		2.3 [1.5; 3.4]	2.3 [1.6; 3.3]	2.5 [1.3; 4.8]	1.1 [0.7; 1.8]	1.6 [1.1; 2.4]	1.7 [1.2; 2.5]	1.6 [1.1; 2.1]
---------------	--	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------	-------------------

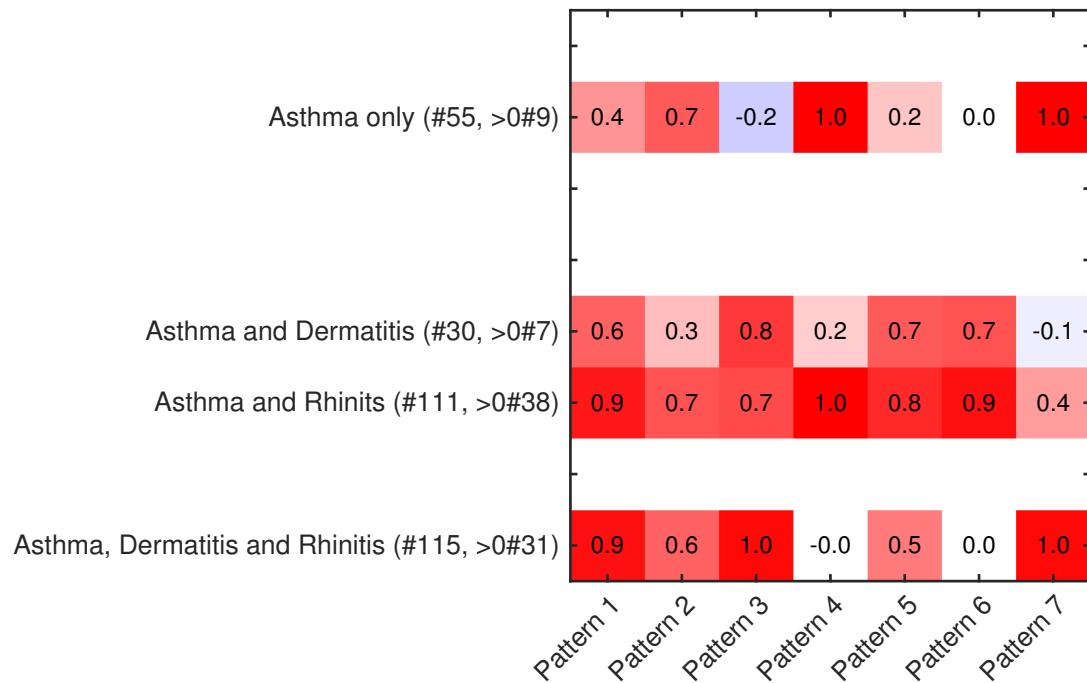
330 § Pattern 1= cat/dog/horse, pattern 2= timothy grass/birch, pattern 3= molds, pattern 4= house dust
331 mites, pattern 5= peanut/wheat flour/mugwort, pattern 6= peanut/soybean, pattern 7= egg/milk/wheat
332 flour
333 * Adjusted for rhinitis 7yrs
334 ^ Adjusted for asthma 7 yrs

Relative Abundance







COPSAC**BAMSE**