A Birth-Death Modelling Framework for Inferring Disease Causality within the Context of Allergy Development

Danielle C. M. Belgrave
Department of Medicine
Imperial College
London SW7 2AZ, UK
Email: d.belgrave@imperial.ac.uk

Konstantina Palla
Department of Statistics
Oxford University
Oxford OX1 3LB, UK
Email: palla@stats.ox.ac.uk

Abstract—**In this paper we deploy a Bayesian non-parametric model to review whether the natural progression of symptoms from eczema to wheeze and rhinitis through chlidhood, often described by the term "allergic march", indicates a causal relationship. Results of our study on real data, suggest that the allergic march may be the result of an ecological fallacy whereby the profiles of eczema, wheeze and rhinitis follow a profile similar to the allergic march when looking at prevalence at a population level. The model based approach recovered distinct classes of profiles of the symptoms, which have distinct clinical associations strongly suggesting the association with distinct pathobiological mechanisms at the individual level. This work is a step towards using model based approaches to understand the aetiology and progression of symptoms and may elucidate early preventative and management strategies aimed towards reducing the global burden of allergic diseases.**

I. Introduction

Understanding and establishing causal mechanisms of diseases remains a grand challenge within the medical domain. Whereas the term "phenotype" refers to the observed manifestations of a disease via its symptoms, "endotype" refers to heterogeneous manifestations of these phenotypes which may be attributed to the distinct underlying pathophysiological and causal mechanisms of a disease which give rise to these observable symptoms. An example of such mechanisms for understanding the cause of disease has been in the domain of asthma and allergic disease with the hypothesized paradigm of the "allergic march" [1] [2]. The allergic march has been hypothesized as a paradigm for inferring causality of allergy during childhood and in later life. Under this hypothesis, children who have eczema in early life (between 0-12 months) have an increased probability of subsequently growing out of eczema and developing asthma (between ages 5-7 years) with the subsequent remission of asthma and development of hay fever (between ages 8-11 years), which then becomes a permanent feature in later life and is labelled as "allergy". This observed profile of cascading symptoms over time has often been suggested as a causal mechanism for explaining the evolution of allergy in early life whereby eczema in early life causes asthma which in turn causes hay fever [3]. Several studies give evidence of this profile at a population level based on cross-sectional snapshots of symptom incidence at different ages across the population. As a consequence of this observed symptom profile, clinical trials have been designed with the aim of targeting the treatment and prevention of eczema in early life in an attempt to stop the subsequent onset of asthma and hay fever in later childhood.

However, the generalized paradigm of the allergic march based on profiles observed on a population level may be indicative of an ecological fallacy; inference about the nature of individuals is erroneously deduced from inference about the general population to which those individuals belong. Under this approach, factors related to each individual, such as a person's genetics and all the environmental influences he or she encounters over a lifetime are not taken into account even though their interactions may act as a distinct underlying causal mechanisms (endotypes) associated with the heterogeneous observed symptoms. Model based approaches arise as a solution towards this, since they provide a flexible framework to define and deploy models that can uncover latent structure in the data that may further give a better insight about the disease and its manifestations. In this work, we deploy the Bayesian non-parametric model by [4] that assigns the patients to latent classes (also called features) over time.

Using the Birth-Death feature allocation process model by [4], the symptoms of eczema, wheeze and rhinitis observed for each patient, depend on their age and corresponding underlying class (feature) assignment. We assume that the underlying structure is an evolving class (feature) allocation of the patients and model this allocation through a birth-death process over time; the process jumps between a state, i.e. a class (feature) allocation, to the next by the addition ('birth') or the 'death' of a class (feature).

Previous studies in machine learning have aimed to elucidate latent features in the healthcare domain [5] [6] [7] . The majority of studies investigating symptom comorbidity has been applied to cross-sectional studies and have employed various clustering algorithms which may be inappropriate for capturing longitudinal within-person heterogeneity over time [8] [9].

These methods are limited in the sense that they assume that feature allocation is fixed over time and do not take into account uncertainty in group membership or longitudinal, evolving characteristics. A limited number of studies have attempted to disaggregate latent profiles of comorbidities, however, they do not necessarily capture evolving feature allocation, which may be a limitation for predicting future events or evaluating feature allocation conditional on previous events, due to the static nature of feature allocation. Within this context, a Bayesian non parametric modelling framework may provide a flexible framework for elucidating dynamic features based on symptom profiles over time [10] [11].

The goal of this work is

- the novel application of a Bayesian (non-parametric) approach to Healthcare Data and,
- the challenge of the long-standing assumption of the "allergic match" through the use of a model based approach.

II. Data Description

Data are taken from the Manchester Asthma and Allergy Study, a population-based birth cohort based in the United Kingdom. The study was approved by local research ethics committees. Informed consent was obtained from all parents; children gave their assent when applicable. Participants were recruited prenatally, and followed prospectively. We used information collected at review clinics at ages one, three, five, eight and 11 years from a total of 712 children who had no missing observations. At each follow-up, validated questionnaires were administered to collect information on parentally-reported symptoms. To assess the presence/absence eczema, asthma and hay fever for each child in the study at each time point, we respectively asked parents the questions:

1) "Has your child in the past 12 months had eczema?"

2) "Has your child had wheezing or whistling in the chest in the last 12 months?"

3) "In the past 12 months, has your child had a problem with sneezing or a runny or blocked nose when he/she did not have a cold or the flu?"

Once the class allocations over time was inferred, we looked for associations of the members of each class to genetic, biological and other factors for which we had information. More specifically, we looked for possible mechanisms which may explain the class assignments and, consequently, the observed symptom profiles. These mechanisms are:

- allergic sensitization: Sensitization is a marker for whether or not someone is allergic to a specific allergen. Data on sensitization ascertained by skin-prick tests at all time points. Sensitization was defined as a wheal diameter 3mm greater than the negative control to at least one allergen out of a panel of common allergens (house dust mite, dog, cat, milk, egg, grass, trees, pollen, peanut, moulds). Sensitization was also assessed as a continuous variable, Immunoglobulin E (IgE), which is a measure of antibody levels produced by the immune system as a reaction to allergens. A higher IgE level is evidence of allergy to a specific allergen,
- the child's biological sex,
- genetic markers known to be commonly associated with eczema (Fillaggrin) and
- lung function measures which measure lung performance through tidal breathing as assessed by Specific Airway Resistance (sRaw) and through forced breathing manoeuvres (such as blowing out a candle) as assessed by Percentage Predicted Forced Expiratory Volume (FEV).

III. Model

We are interested in time series settings where we observe data $\{Y_t \in \mathcal{Y} : t = 1, \dots, L\}$. More specifically, we consider problems where the observations are explained by a latent structure which assigns objects to features (classes) and this feature allocation changes over time. We use the birth-death feature allocation process (BDFP) first presented in the paper by [4]. The process is a Markov Jump process where the events are the birth and the death of a feature. To facilitate inference, we use the finite construction of the process, the Beta Event process (BEP) which we describe in what follows.

A. The Beta Event Process

We consider the finite approximation of the BDFP [4] which gives the countably infinite model in the limit. We consider a nonhomogenous Poisson process $\mathbf{\Pi}$, on the space $\mathbb{S} = [0,1] \otimes \mathbb{X} \otimes [0,T] \otimes [0,\infty)$, with the Lévy measure $\nu(\mathrm{d}\omega \mathrm{d}x \mathrm{d}t_b \mathrm{d}t_\omega)$. A sample $\mathbf{\Pi} = \{\omega_k, x_k, t^k_b, t^k_\omega\}_k$ corresponds to a set of atom with $k = 1, \dots |\mathcal{F}|$. Each atom corresponds to a feature and is associated with a weight $\omega_k \in [0,1]$, a location x_k, a birth time $t^k_b \in [0,T]$ and a life-span $t^k_\omega \in [0,\infty)$. We restrict the space of t_b to be $[0,T]$ instead of the whole real line $\mathbb{R}$. This accounts for typical applications of the model where we observe data at distinct times over a finite time range.

The process is depicted in Figure 1 and the infinite case can be derived as the limit $K \to \infty$ of the following:

- Consider a time range $[0,T]$ and a set of features $\mathcal{F}$, such that $|\mathcal{F}| \sim \text{Poisson}(KT)$. Assign to each feature $f_k \in \mathcal{F},\ k = 1, \dots |\mathcal{F}|$ a weight ω, such that $\omega_k \sim \text{Beta}\big(\frac{R}{K}, 1\big)$ and $\mathbf{\Omega} = [\omega_1, \omega_2 \dots \omega_{|\mathcal{F}|}]$
- Associate each feature $f_k \in \mathcal{F},\ k = 1, \dots |\mathcal{F}|$ with a birth time t^k_b uniformly sampled in $[0,T]$; $\quad t^k_b \sim \mathcal{U}(0,T)$ and $\mathbf{t}_b = [t^1_b \dots t^{|\mathcal{F}|}_b]$.
- For each $f_k \in \mathcal{F}$, sample its life span $t^k_w \sim \text{Exponential}(D)$, where $D = \frac{R}{\alpha}$ is the death rate. Define the time of death t^k_d as $t^k_d = t^k_b + t^k_w$ and $\mathbf{t}_w = [t^1_w \dots t^{|\mathcal{F}|}_w]$.

We call the sequence of the above steps **Beta Event Process (BEP)**. Putting everything together, generate a sample $B = \{\mathcal{F}, \mathbf{\Omega}, \mathbf{t}_b, \mathbf{t}_w\} \sim \text{BEP}(\alpha, R, K, T)$ as follows:

$$|\mathcal{F}| \sim \text{Poisson}(KT),$$
$$\omega_k \sim \text{Beta}\left(\frac{R}{K}, 1\right), \quad t^k_b \sim \mathcal{U}(0,T), \quad t^k_\omega \sim \text{Exponential}(D) \tag{1}$$

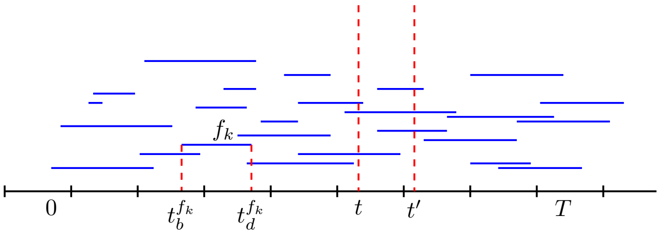

Fig. 1: Cartoon for the Beta event process: A $\text{Poisson}(KT)$ number of features are uniformly distributed across the time range $[0, T]$ (blue lines). Each feature is assigned a weight sampled from $\text{Beta}(\frac{R}{K}, 1)$. The leftmost point of each line corresponds to the time of birth of that feature, while the length of each line indicates the life span of each feature sampled from $\text{Exponential}(D)$. To sample feature allocations from the process, we consider random time points across time, e.g. t, t' and draw imaginary red lines. The feature allocation matrix at t involves the features that are crossing the red line at t. The membership of the objects $n = 1, \ldots, N$ to those features is defined by the values of the corresponding elements in the potential matrix S.

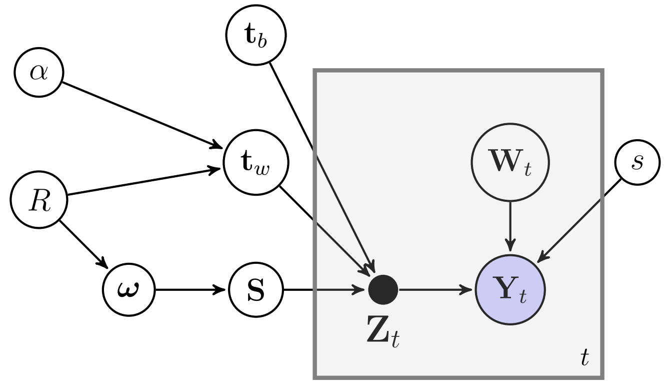

Fig. 2: Graphical representation of the model for a time point t and for sigmoid likelihood. The time series $\mathbf{Z}$ and $\mathbf{Y}$ are represented as single nodes indexed by the time location t. The birth and life span times of the total KT features are depicted using vector notation $\mathbf{t}_b$ and $\mathbf{t}_w$. The black ($\mathbf{Z}_t$) and grey ($\mathbf{Y}_t$) nodes indicate deterministic and observed parameters respectively.

for $k = 1, \ldots, |\mathcal{F}|$. Having drawn a sample B from the BEP, we can construct the feature allocations over time as follows

$$S_{nk}|\omega_k \sim \text{Bernoulli}(\omega_k), \quad Z_{nk}(t) = S_{nk}\mathbb{I}(t_b^k < t < t_b^k + t_\omega^k) \tag{2}$$

where $n = 1, \ldots, N$. The binary matrix $\mathbf{S}$ of dimension $N \times |\mathcal{F}|$ is a *feature potential* matrix. Each binary element S_{nk} indicates whether object n possesses feature f_k. S is a global variable and doesn't depend on time t. At any time t, the feature allocation matrix $Z(t)$ is a deterministic function of the current features present at t, that is $\{f_k : t_b^k < t < t_b^k + t_w^k, k = 1, \ldots, |\mathcal{F}|\}$ and the feature potential matrix S, i.e. $Z_{nk}(t) = 1$ iff $S_{nk} = 1$ and $t_b^k < t < t_b^k + t_\omega^k$. The resulting feature allocation process $(z_n(t))_{\mathbb{T}}$ is equivalent to the following: every time a new feature f_k is created, each object n joins with probability ω_k, i.e. $z_{nk}(t_b^k)|\omega_k \sim \text{Bernoulli}(\omega_k)$. If $z_{nk}(t_b^k) = 1$, object n will possess feature f_k until $t_b^k + t_\omega^k$. Repeat this process for all objects. Moreover, each $Z(t)$ for $t \in \mathbb{T}$ is a matrix of dimensions $N \times F^{(t)}$ and $F^{(t)} \leq |\mathcal{F}|$. Figure 2(a) show the graphical model for the BEP and for the sigmoid likelihood used in this paper.

a) Hyperpriors.: We put gamma priors on α and R.

b) Likelihood model.: Let Y_t be the $N \times D$ binary matrix that relates children to symptoms at time point t , i.e. $y_{tnd} = \mathbf{Y}_t(n, d) = 1$ iff the n-th child has symptom d at time t. and 0 otherwise. The symptoms here are $D = 3$; eczema, wheeze and rhinitis. The probability of a child having a symptom at any particular time point is determined by the combined effect of all the features present at that time. Let $\mathbf{W}_t$ be a $|\mathcal{F}| \times D$ real-valued weight matrix where $W_t(k, d)$ is the weight that affects the probability of the nth child having symptom d if the child has feature k on, i.e. $Z_{tnk} = Z_t(n, k) = 1$. The observations are independent conditioned on $\mathbf{Z}_t$ and $\mathbf{W}_t$, and only the features that are on for the nth child at time t influence the probability at that time (see Figure 2). Formally,

$$P(y_{tnd} = 1|\mathbf{Z}_t, \mathbf{W}_t) = \sigma\Big(\sum_k Z_{tnk} W_{tkd} + s\Big) \tag{3}$$

for $k = 1, \ldots, |\mathcal{F}|$, where s is a bias term and $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. For completeness, we assume the priors $w_t(k, l) \sim \mathcal{N}(\mu_w, \sigma_w^2)$ and $s \sim \mathcal{N}(\mu_s, \sigma_s^2)$. In the experiments assumed $\mu_s = 0, \sigma_s = 1$ and $\mu_w = 1, \sigma_w = 1$.

B. Inference

We employed Markov Chain monte Carlo (MCMC) for posterior inference over the latent variables of the model. A detailed description of the sampling steps is provided in the supplementary material.

IV. Results

We created 3 train-test splits holding out 20% of the data, and ran 2500 MCMC iterations with 2K burn-in. We also thresholded the possible number of classes inferred by the BEP to 12. This choice was based on initial experiments that showed that this thresholded was enough to capture all the classes present in the data. To elaborate over the use of time evolving feature allocations as opposed to static ones in the discovery of subtypes of complex diseases, we compared the temporal BEP to independent models at each time point, that is static feature allocations as found by the Indian Buffet process [12, IBP] when applied to each time point (age) independently. We see that in terms of predictive performance the dynamic model outperforms the independent IBP models (Table I). Both models have comparable performance in test error. However, the BEP claims a statistically significant performance in test likelihood.

TABLE I: Dataset results using 20% held out data, a truncation level of $|\mathcal{F}| = 12$, 2500 iterations and a burn-in of 2000. Results are the average over 3 MCMC chains.

	BEP	independent IBP
Train error	$\mathbf{0.1945 \pm 0.0313}$	0.2677 ± 0.0244
Test error	$\mathbf{0.2546 \pm 0.0381}$	0.2660 ± 0.0277
Train log likelihood	$\mathbf{-2.683 \pm 0.0340}$	-3.4019 ± 0.0289
Test log likelihood	$\mathbf{-1.0540 \pm 0.01567}$	-1.2390 ± 0.0159

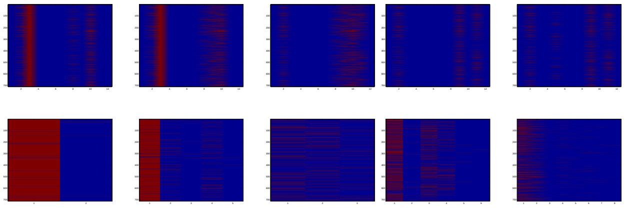

Fig. 3: Inferred feature allocation matrices for the five time points (from left to right) in the dataset. **First row:** Feature allocation matrices inferred by BEP. **Second row:** Feature allocation matrices inferred by independent IBP. Red and blue indicate membership and non-membership respectively.

Looking at Figure 3, the two models provide a different picture of the allocation. The BEP model identified 7 distinct classes in total over all the ages. The first rowshows a relatively slow evolution of the class assignments (feature allocations). This is also confirmed in the Figure 1 of the supplementary paper. Moreover, we notice that for age 1 and age 3 the inferred feature allocations are more similar than the ones between the age 5 and 8. This underlines the fact that the BEP allows for class assignments to be similar the closer they are in time. The model inferred a class (feature) death rate ≈ 0.3 which corresponds to increased life time for each class and thus allowing for slow change in the class assignments. In the IBP model, we identified 2 classes at age 1, 5 classes at age 3, 3 classes at age 5, 6 classes at age 8 and 8 classes at age 11 years. These classes are shown in the low panel of Figure 3, and because of lack of evolution in the classes and reasons previously stated, we do not illustrate further analysis.

Both models seem to recover the classes with the largest membership. However, the BEP is able to uncover additional classes with smaller -but still- significant membership. The IBP constrains the allocation mainly to classes with considerable membership. Analysis (not shown) also showed that since the IBP model was based on classes which were determined statically at a given time point, they were poorer at distinguishing between clinical outcomes compared to the BEP where different classes had different patterns of association, indicating the relevance of a larger number of classes which evolve over time.

In Figure 4 we present, for each symptom, the proportion of the children that manifest it out of the total children assigned in each recovered class. The bottom panel shows the distribution of eczema, wheeze and rhinitis over time for classes identified in the BEP model. The BEP model identified 7 classes with distinct profiles of eczema, wheeze and rhinitis over time, and we assigned an additional class of children who were not assigned to any class at any time point. Class 2 had a profile of a high probability of eczema, wheeze and rhinitis at all time points. Class 3 had a high probability of wheeze and eczema in early life and a low probability of rhinitis in early life, with subsequent dissolution of all symptoms. Class 5 only appear at age 11, with a high probability of eczema, wheeze and rhinitis at this time point. Class 8 has resolution of symptoms by age 5 and is characterized by having the highest probability of wheeze in early life. Class 9 starts manifesting symptoms at age 3 years, with a high probability of eczema and rhinitis and a low probability of wheeze. Class 11 starts manifesting symptoms at age 5 years, and has the highest probability of rhinitis, with moderate eczema and wheeze. Children labelled as "No Class Allocation" had a low probability of symptoms throughout life. A similar analysis is hard to achieve in the static IBP case due to non-identifiability of the distinct classes over time.

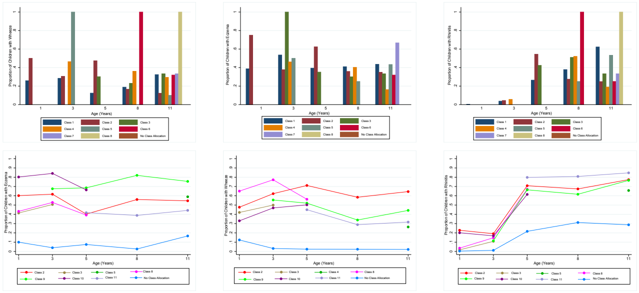

Fig. 4: Proportion of children experiencing eczema, wheeze and rhinitis by class membership **First row:** Proportions for the IBP model. **Second row:** Proportions for the BEP model

To get a qualitative insight into the classes recovered by the model, we looked for association of the class members with certain biological processes and genetic markers, for which we had side information for related to each patient. Where class membership only has an effect on the average profile of the symptom, but not on the evolution of the symptom over time, the average variation over time with the class was omitted. Statistically significant results as indicated by the 95% Lower and Upper Confidence Bounds (LCB and UCB). Results highlighted in bold indicate that the class is significantly associated with the outcome in terms of the Confidence Bounds. In Table II, we used longitudinal regression models, the output is the lung function value in the range $(0, 3)$, predictor is age and class membership and the parameter estimates are the weights for class membership. Table III shows the association of the class members with Fillaggrin, a known genetic variant of asthma and allergic diseases which may distinguish severity of disease. Here, we used a logistic regression model with the class label as the output and the fillagrin indicator 0/1 value as the predictor. Table V was constructed as Table II, but it includes an interaction term between class membership and time/age. Class 2 was distinguished by poor lung capacity, as indicated by higher specific airway resistance and lower Forced Expiratory volume. These children also had a higher probability of sensitization which increased over time. Both Class 9 and 10 are associated with Fillaggrin, a genetic mutation which is known to increase the risk of allergy, however, children in class 9 have a high probability of sensitization in early life which decreases over time, whereas class 10 starts off with a comparatively low probability of sensitization to mite which increases over time. Children with "No Class Allocation" were significantly less likely to have the Fillaggrin genetic mutation and had better lung function and lower risk of sensitization to allergens. Once again, such an analysis in the independent IBP model was not facilitated since there was no stable pattern of associations over time.

TABLE II: Results for specific airway resistance

	Specific Airway Resistance			
	Parameter Estimate	Standard Error	LCB	UCB
Class 2	**0.070**	**0.033**	**0.004**	**0.135**
Class 3	-0.026	0.015	-0.056	0.004
Class 5	0.004	0.035	-0.064	0.072
Class 8	0.029	0.024	-0.017	0.075
Class 9	0.019	0.017	-0.014	0.052
Class 10	0.010	0.016	-0.021	0.040
Class 11	-0.016	0.017	-0.049	0.017
No Class Allocation	**−0.050**	**0.016**	**−0.079**	**−0.017**

TABLE III: Results for Fillaggrin Genetic Mutation

	Fillaggrin Genetic Mutation			
	Parameter Estimate	Standard Error	LCB	UCB
Class 2	0.368	0.310	-0.240	0.975
Class 5	0.443	0.408	-0.358	1.243
Class 8	-0.070	0.423	-0.890	0.758
Class 9	**0.708**	**0.275**	**0.168**	**1.247**
Class 10	**0.607**	**0.270**	**0.078**	**1.135**
Class 11	0.002	0.321	-0.629	0.630
No Class Allocation	**−0.838**	**0.334**	**−1.492**	**−0.183**

TABLE IV: Results for sensitization to any allergen

	Sensitization to Any Allergen				**Sensitization to Any Allergen Variation Per Year**			
	Parameter Estimate	Standard Error	LCB	UCB	Parameter Estimate	Standard Error	LCB	UCB
Class 2	**1.599**	**0.464**	**0.689**	**2.509**	**0.104**	**0.049**	**0.009**	**0.199**
Class 3	**−2.834**	**0.452**	**−3.721**	**−1.948**	**0.744**	**0.141**	**0.467**	**1.020**
Class 5	**−1.088**	**0.511**	**−2.090**	**−0.087**				
Class 8	**−1.674**	**0.717**	**−3.079**	**−0.269**	**0.487**	**0.168**	**0.158**	**0.816**
Class 9	**2.855**	**0.402**	**2.069**	**3.641**	**−0.045**		**−0.178**	**−0.003**
Class 10	**1.512**	**0.234**	**1.054**	**1.972**				
Class 11	**1.927**	**0.277**	**1.385**	**2.469**				
No Class Allocation	**−2.208**	**0.292**	**−2.779**	**−1.636**				

TABLE V: Results for sensitization to Mite

	Sensitization to Mite				**Sensitization to Mite Variation Per Year**			
	Parameter Estimate	Standard Error	LCB	UCB	Parameter Estimate	Standard Error	LCB	UCB
Class 2	**2.308**	**0.404**	**1.516**	**3.099**				
Class 3	**−7.107**	**1.031**	**−9.127**	**−5.086**	**2.082**	**0.336**	**1.423**	**2.740**
Class 5	-1.058	0.617	-2.267	0.151				
Class 8	**−5.227**	**1.597**	**−8.357**	**−2.097**	**0.982**	**0.345**	**0.306**	**1.658**
Class 9	**3.679**	**0.5**	**2.698**	**4.659**	**−0.167**	**0.053**	**−0.272**	**−0.062**
Class 10	**−1.938**	**0.655**	**−3.222**	**−0.655**	**0.668**	**0.141**	**0.392**	**0.945**
Class 11	**2.991**	**0.664**	**1.689**	**4.293**	**−0.184**	**0.078**	**−0.336**	**−0.031**
No Class Allocation	**−2.572**	**0.413**	**−3.381**	**−1.762**				

V. Conclusion

In this study, we were able to identify distinct classes of profiles of eczema, wheeze and rhinitis, which have distinct clinical associations and may be endotypes of allergic disease. Two of these classes were strongly associated with a genetic marker, Fillaggrin, which is commonly identified with allergy severity. This indicates that this genetic marker may be causally linked to some manifestations of symptoms over time and may thus elucidate a plausible underlying biological mechanism for these endotypes. We also identified time-varying different patterns of allergic and lung function development for different classes. This separation of distinct profiles with distinct associated factors indicates that the “allergic march” hypothesis may be an inadequate description of the progression of symptoms on an individual level.

The birth-death latent feature allocation framework we described in this paper may be generalizable to other disease areas which are represented by comorbidity of symptoms. Such models may help elucidate a more robust clinical framework for understanding disease heterogeneity and therefore may elucidate distinct underlying causal mechanisms of different profiles of symptom co-occurrence, leading to targeted and personalised disease treatment, management and intervention strategies. Identification of possible causal mechanisms of distinct classes may allow pharmaceutical companies to develop more targeted therapies. The presented work falls in the more general class of model based approaches and underlines their efficiency throughout understanding diseases.

We have focused on children for whom all data at all time points was fully observed. Future work will look into methods of compensating for missing data in scenarios where data may not be missing at random or where we want to infer whether or not there is a plausible mechanism that may explain missing data.

Acknowledgment

Konstantina’s research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

References

[1] U Wahn. What drives the allergic march? *Allergy*, 55(7):591–599, 2000.

[2] Bruce R Gordon. The allergic march: can we prevent allergies and asthma? *Otolaryngologic clinics of North America*, 44(3):765–777, 2011.

[3] Jonathan M Spergel and Amy S Paller. Atopic dermatitis and the atopic march. *Journal of Allergy and Clinical Immunology*, 112(6):S118–S127, 2003.

[4] Konstantina Palla, David A. Knowles, and Zoubin Ghahramani. A birth-death process for feature allocation. Sydney, Australia, August 2017.

[5] Anne M Fitzpatrick, W Gerald Teague, Deborah A Meyers, Stephen P Peters, Xingnan Li, Huashi Li, Sally E Wenzel, Shean Aujla, Mario Castro, Leonard B Bacharier, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the national institutes of health/national heart, lung, and blood institute severe asthma research program. *Journal of allergy and clinical immunology*, 127(2):382–389, 2011.

[6] Tamazoust Guiddir, Philippe Saint-Pierre, Elsa Purenne-Denis, Nathalie Lambert, Yacine Laoudi, Rémy Couderc, Rahelé Gouvis-Echraghi, Flore Amat, and Jocelyne Just. Neutrophilic steroid-refractory recurrent wheeze and eosinophilic steroid-refractory asthma in children. *The Journal of Allergy and Clinical Immunology: In Practice*, 2017.

[7] Wenbin Zhang, Jian Tang, and Nuo Wang. Using the machine learning approach to predict patient survival from high-dimensional survival data. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 1234–1238. IEEE, 2016.

[8] J Garcia-Aymerich, M Benet, Yvan Saeys, M Pinart, X Basagana, HA Smit, V Siroux, J Just, I Momas, F Rancière, et al. Phenotyping asthma, rhinitis and eczema in medall population-based birth cohorts: an allergic comorbidity cluster. *Allergy*, 70(8):973–984, 2015.

[9] N Ballardini, A Bergström, C-F Wahlgren, M Hage, E Hallner, I Kull, E Melén, JM Antó, J Bousquet, and M Wickman. Ige antibodies in relation to prevalence and multimorbidity of eczema, asthma, and rhinitis from birth to adolescence. *Allergy*, 71(3):342–349, 2016.

[10] Yanxun Xu, Peter Müller, and Donatello Telesca. Bayesian inference for latent biologic structure with determinantal point processes (dpp). *Biometrics*, 2016.

[11] Peter J Green. Mad-bayes matching and alignment for labelled and unlabelled configurations. *Geometry Driven Statistics*, 121:377, 2015.

[12] Thomas L. Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, July 2011.